QoS-aware and Policy Based Mobile Data O oading

Amani, Mojdeh

Awarding institution:
King's College London

QoS-aware and Policy Based Mobile Data Offloading

Mojdeh Amani

Centre for Telecommunications Research

King's College London

A thesis submitted to King's College London in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

September 2013

To my loving parents, and my brother *Mehran*.

Acknowledgements

First and foremost, I would like to thank my Lord who gave me the strength and the means to embark upon the journey to my PhD.

This thesis would not have been possible without the guidance and the support of several individuals who helped in making my time at CTR one of the most memorable experiences of my life. I am deeply indebted to my supervisor Prof. Hamid Aghvami, one of the most amazing personalities who supported me all the way in this journey through the happier and the non-happier times. I will always be thankful to him and respect him immensely. He provided a stimulating environment at CTR and his motivation, knowledge and long lasting experience have been invaluable to me. I would like also to thank Dr. Toktam Mahmoodi, who supported me throughout this journey. Not only did she teach me how to conduct quality research, she also taught me to maintain my integrity as a researcher. She is one of the most committed researcher who I have ever worked with. I will never forget the long hours that she spent helping me with my work. Indeed, she is a respectable researcher for CTR.

I am particularly grateful to my colleagues at CTR who gave me hours of enjoyable company. Dev and George who I have spent most of my first year with, are among the friends who I never forget. Merat and Nika were always the friendliest face of CTR with whom you do not need to pretend being somebody else; you are accepted as you are. Reza whom his company during these years specially writing up period is among the most unforgettable memories of this journey. Especially, I would like to acknowledge the efforts of Adnan who always gave me useful tips related to my work and Amir who spend long hours to help me for my research. I have enjoyed the company of many of my colleagues and could make lots of valuable friends at CTR. My thanks for the rest of my colleagues at CTR, (in no particular order) go to Yaqub, Alex, Panagiotis, Diogo, Mona, Hossein, Mohammad, Hadi, Nur, Ana, Arman and Saba for providing such a great memory.

Outside CTR, I wish to thank all my friends who supported me whenever needed. Among all are Yalda and Leila whom are my friends forever and Maryam who is a very sweet accompany and make my time in in London very enjoyable.

My last but not least thoughts go to my parents. I cannot find words to do justice to my Mom and Dad. My parents have always provided unconditional support and appreciation for my scholarly endeavors. They taught me how to aim high in life and succeed with true devotion. With no doubt I can say they are the only ones who understand me and love me in all circumstances. My brother, Mehran, whom I love him the most, and whom his appreciation and kindness have always been the very strong reasons for me to experience and explore new things in life. I would have loved to be a good role model as her older sister and I hope I could. I dedicate my thesis to all my beloved ones, either named here or not.

Abstract

The rapid growth in the number of 3G/4G enabled devices such as smartphones and tablets has created exceptional demand for ubiquitous connectivity and high quality applications. As a result, cellular networks are struggling to keep up with this explosive demand for data traffic. The emergence to LTE has boosted cellular network throughput; however these improvements are not sufficient given the limited availability of licensed spectrum. To meet the requirements of capacity-hungry applications, Wi-Fi offloading has been intensively researched as an essential approach to alleviate the mobile data traffic load on cellular network by providing extra capacity and improving overall performance. The offloading algorithms should be evaluated and compared to steer Wi-Fi offloading to increase the combined throughput and network performance of LTE and Wi-Fi access technologies connected to the evolved packet core (EPC) with at least the baseline case of having all the data traffic in LTE. In this thesis, novel offloading algorithms are proposed and implemented to address challenges in Wi-Fi offloading to LTE networks and provide solutions when performance needs exceed the capability of the LTE access network. In the design of such smart offloading techniques, important issues such as scalability and stability are being considered. Through an extensive set of simulations, the performance of the proposed techniques is thoroughly investigated focusing on the figure of merits that affects user experience. The end-to-end throughput that a flow can accomplish, offloading efficiency and packet dropping rate are examined. Furthermore, these evaluations have demonstrated that offloading users from LTE to Wi-Fi reduces burden on the LTE network without affecting user experience. Also it is shown that the mobile communication architecture can be improved further by applying the principles of Software-Defined Networking (SDN) with providing logically centralized control of the overall infrastructure, and enabling programmability.

Contents

Contents **6**

List of Figures **7**

List of Tables **8**

List of Acronyms **9**

1 Introduction **16**
1.1 Scope of the Work 17
1.2 Thesis Contributions 18
1.3 Thesis Outline 19

2 Background Study **21**
2.1 Today's Mobile Network Challenges 22
2.2 Evolution of Wireless Access Networks 23
2.2.1 UMTS 24
2.2.2 LTE 25
2.2.3 LTE-Advanced 25
2.2.4 HetNet 26
2.2.5 Wireless Local Area Network (WLAN) and Wi-Fi 26
2.2.6 Worldwide inter-operability for Microwave Access (WiMAX) . . 26
2.2.7 Software Defined Networking 27
2.2.8 SDN for Cellular Data Networks 27
2.2.9 Cloud Computing 29
2.3 Offloading & Offloading Techniques 29
2.3.1 Mobile Data Offloading via Wi-Fi 29
2.3.2 Mobile Data Offloading via Femtocells 31
2.3.3 Advantages and Disadvantages of Wi-Fi and Femtocell 32

2.3.4 Mobile Data offloading via Wimax . . . 32
2.3.5 Mobile data Offloading via Opportunistic Communication . . . 33
2.3.6 Mobile Data Offloading via IP Flow Mobility . . . 33
2.3.7 Core network Offloading . . . 34
2.3.8 Media Optimization Solutions . . . 34
2.3.9 Challenges of Mobile Data Offloading . . . 35
2.4 Quality of Service (QoS) in Mobile Networks . . . 36
2.4.1 Policy Management . . . 37
2.4.2 QoS and Policy Management in LTE and EPC Networks . . . 37
2.4.3 Policy Decision Definition . . . 40
2.4.4 Centralized Policy Control . . . 41
2.4.5 Distributed Policy Control . . . 41
2.4.6 Functional Elements in Implementing Policy and QoS in EPC . 41

3 QoS-Aware Mobile Data Offloading 43
3.1 Background Study . . . 45
3.1.1 LTE Cellular Network . . . 45
3.1.2 Seamless Wi-Fi Offloading . . . 47
3.1.3 IP Flow Mobility . . . 48
3.1.4 Establish IP Flow with Service Continuity . . . 48
3.1.5 IP Flow Mobility Scenarios . . . 49
3.1.6 LTE MAC Scheduler . . . 51
3.1.7 Related Works . . . 52
3.2 QoS-Aware MAC Scheduler and Offloading . . . 54
3.3 Problem Statement . . . 54
3.3.1 QoS-Aware Flow Scheduling . . . 55
3.3.2 Offloading Mechanism . . . 57
3.4 Performance Evaluation . . . 58
3.4.1 Simulation Scenario . . . 59
3.4.2 System Model for LTE . . . 60
3.4.3 Simulation Methodology . . . 61
3.4.4 Numerical results . . . 62
3.5 Conclusion . . . 62

4 Policy-Based Mobile Data Offloading 67
4.1 Introduction . . . 67
4.2 Policy Based Offloading Framework . . . 68

4.2.1 User Centric Offloading Policies . . . 68
4.2.2 Network Centric Offloading Policies . . . 70
4.2.3 Hybrid Offloading Polices . . . 70
4.3 Performance Evaluation . . . 74
4.3.1 System Model . . . 74
4.3.2 Traffic Models . . . 74
4.3.3 Simulation Methodology . . . 76
4.3.4 Numerical Results . . . 77
4.4 Conclusion . . . 78

5 Programmable Policies for Mobile Data Offloading 80
5.1 LTE and Wi-Fi Interworking Architecture . . . 82
5.2 System Model . . . 85
5.2.1 Policy Control . . . 85
5.2.2 SDN-Controller and Mobile Data Offloading . . . 87
5.3 Policy Derivation and Offloading Mechanism . . . 88
5.3.1 Policy Derivation . . . 88
5.3.2 Offloading Mechanism . . . 90
5.4 performance Evaluation . . . 92
5.4.1 Simulation Model . . . 92
5.4.2 Simulation Scenarios and Numerical Results . . . 94
5.5 Concluding Marks . . . 96

6 Conclusions and Future Research 99
6.1 Avenues of Future Research . . . 100
6.1.1 Opportunistic small cells and 3G/4GWi-Fi interworking . . . 101
6.1.2 Offloading Decision in a Framework of HetNet and Mobile Cloud Computing . . . 101
6.1.3 Software Defined Cloud Networking (SDCN) . . . 101

References 102

List of Figures

2.1 Mobile Data offloading via Wi-Fi 31
2.2 Mobile Data offloading via Femtocell 31
2.3 Mobile Data offloading via Opportunistic Communication 33
2.4 Mobile Data Offloading Via IP Flow Mobility 34
2.5 Mobile Data Offloading Via Core Network 34

3.1 LTE Resource Block 47
3.2 IP flow Mobility 49
3.3 eNodeB Scheduler 52
3.4 LTE Resource Block 55
3.5 Aggregate Voice Throughput 63
3.6 Aggregate Video Throughput 63
3.7 Aggregate Data Throughput 64
3.8 Number of Completed Voice Flows within Delay Requirement 64
3.9 Number of Completed Video Flows within Delay Requirement 65
3.10 Number of Completed Data Flows 65

4.1 Policy Based Offloading and Mechanism(step1,2,3 correspond to Monitoring, Decision Making and Execution respectively 72
4.2 Cellular Busy Hour Load 75
4.3 Wi-Fi User Association Rate 75
4.4 Offloading Efficiency over a 24 hour period against the BusyLoad. . . . 78
4.5 Blocking ratio over a 24 hour period against the BusyLoad. 79

5.1 Non-Roaming 3GPP Arch. for Non-3GPP IP Access Integration into EPC using S5, S2a, S2b and S2c 83
5.2 Trusted non-3GPP IP Access Integration into EPC 84
5.3 Coupling of RRM and NRM to optimize utilization of both radio and network resources 86

5.4 Coupling RRM and NRM via SDN-controller in the non-3GPP IP access integration into the EPC 89
5.5 SDN enabled offloading procedure 91
5.6 Offloaded Traffic Rate, and Dropping Rate in percentage Vs No. of Flows for Online and Offline Policy Derivation Methods. 95
5.7 No. of Offloaded and Dropped Flows vs No. of Flows for Online and Offline Policy Derivation Method 96
5.8 Offloaded Traffic Rate, and Dropping Rate in percentage Vs No. of Flows for Online and Offline Policy Derivation Methods. 97
5.9 No. of Offloaded and Dropped Flows vs No. of Flows for Online and Offline Policy Derivation Method 97

List of Tables

2.1 Standard LTE QCIs 39

3.1 Performance Requirement by Service Category 60
3.2 Simulation Parameters 61

4.1 Traffic Model Parameters 76

5.1 Performance Requirement by Service Category 93

List of Acronyms

1G First Generation Mobile Networks

2G Second Generation Mobile Networks

3G Third Generation Mobile Networks

3GPP Third Generation Partnership Project

4G Fourth Generation Mobile Networks

AAA Authentication, Authorization and Accounting

AMBR Aggregate Maximum Bit Rate

AN Access Network

AP Access Point

APN Access Point Name

ARP Allocation and Retention Priority

BER Bit Error Rate

BPSK Binary Phase Shift Keyin

BS Base Station

BSS Base Station Subsystem

CAGR Compound Annual Growth Rate

CAPEX Capital Expense

CDMA Code Division Multiple Access

DPI Deep Packet Investigation

DSMIPv6 Dual Stack Mobile IPv6

EDGE Enhanced Data Rate for GSM Evolution

eNodeB evolved NodeB

EPC Evolved Packet Core

EPS Evolved Packet System

ETSI European Telecommunications Standards Institute

E-UTRAN Evolved-Universal Mobile Telecommunications System Terrestrial Radio Access

FDD Frequency Division Duplex

FDMA Frequency Division Multiple Access

GBR Guaranteed Bit Rate

GGSN Gateway GPRS Support Node

GPRS General Packet Radio Subsystem

GSM Global System for Mobile Communications

HetNet Heterogenous Network

HLR Home Location Register

HSS Home Subscriber Server

IETF Internet Engineering Task Force

LTE Long Term Evolution

LTE-A LTE Advanced

LIPA Local IP Access

MAC Medium Access Control

MBR Maximum Bit Rate

MIMO Multiple-Input and Multiple-Output

MISO Multiple-Input and Single-Output

MPLS Multi-protocol Label Switching

MS Mobile Station

non-GBR non-Guaranteed Bit Rate

OFDM Orthogonal Frequency Division Multiplexing

OFDMA Orthogonal Frequency Division Multiple Access

SC-OFDMA Single Carrier-FDMA

PCC Policy Control and Charging

PCEF Policy and Charging Enforcement Function

PCRF Policy and Charging Rules Function

PDN-GW Packet Data Network Gateway

PHY Physical

QCI QoS Class Indicator

QoS Quality of Service

QPSK Quadrature Phase Shift Keying

RF Radio Frequency

RNC Radio Network Controller

SDF Service Data Flows

SDN Software Defined Networking

SINR Signal to Interference plus Noise Ratio

SIPTO Selected IP traffic Offload

SNR Signal to Noise Ratio

TACS Total Access Communication System

TDMA Time Division Multiple Access

TFT Traffic Flow Templates

TCP Transport Control Protocol

UE User Equipment

UDP User Datagram Protocol

UMTS Universal Mobile Telecommunications System

WCDMA Wide-band Code Division Multiple Access

WiFi Wireless-Fidelity

WiMax Worldwide inter-operability for Microwave access

WMN Wireless Mesh Network

WLAN Wireless Local Area Network

Chapter 1

Introduction

According to Cisco forecasts, the annual global IP traffic will reach the zettabyte threshold by the end of 2017. Global IP traffic has increased eight times over the past 5 years, and will increase four times more over the next 5 years. Overall, IP traffic will grow at a compound annual growth rate (CAGR) of 32 percent from 2010 to 2017 [1]. Mobile internet access is becoming more and more popular today and the ubiquitous cellular data networks attract millions of users, which in turn creates pressure on the limited spectrum of these networks; consequently, subscribers in big cities are suffering from poor quality as the networks struggle to provide the required service. Over the last few years, a variety of access technologies has been deployed. While the 2G cellular systems have evolved into 3G systems, such as UMTS, providing wide area coverage, Long Term Evolution facilitate the usage of wider spectrums and complement 3G networks to increase capacity and handle more traffic. However, the cellular networks do not have enough capacity to accommodate the ever growing data traffic. Therefore, the limitation of radio link improvement and the spectrum available for the operators, makes it imperative to seek other solutions.

In response to the high demand for mobile internet, different solutions have been considered to relieve the congestion. The only viable solution for increasing overall mobile network capacity is to increase the carrier-to-interference ratio while decreasing the cell size and utilizing a common core network to deploy small cell technologies. The coexistence of different wireless technologies in the same area could overcome some of the most common issues of wireless access networks, such as limited bandwidth, incomplete coverage and time-varying channel conditions. Users equipped with multiple interfaces could benefit from the highest possible quality of service, unaware of the employed access technology. Therefore, enabling the terminal to switch from one access network to another could optimize performance and resource utilization while the

service continuity would be guaranteed in a transparent way. In order to achieve this goal, new solutions and technologies need to be developed to provide the users with optimum performance.

Mobile data offloading or data offloading is considered as a key solution both by mobile operators and vendors as the data traffic on mobile networks continues to grow rapidly and is associated with congestion and degradation of user experience. Most of the mobile operators have started to consider and implement mobile data offloading techniques. In an evaluation of each technology's strength, it will be clear that small cells, along with Wi-Fi, will empower operators to handle the capacity challenge of future networks. This will result in increased development of small cell access points, which combine technologies to benefit from both licensed and unlicensed spectrum technical advantages and use all the available spectrums to challenge the growth of data traffic. Although there are many wireless access technologies interworking architectures available in the literature, the quality of service guarantee and seamless roaming are two major drawbacks in these proposed architectures.

In order to provide a seamless user experience in heterogeneous networks, a more consistent traffic prioritization and core network policy enforcement, leading to intelligent offloading decisions which consider QoS parameters and the availability of service and access networks, would be preferable. To overcome these challenges, a smart offloading mechanism is needed to improve the data traffic management and network selection, based on network condition, user application and QoS parameters.
This challenging issue has inspired a large body of research over the last few years, of which an overview is given in this thesis.

1.1 Scope of the Work

This thesis investigates the performance improvement of a heterogeneous network which consists of high power macro-cells (LTE) and low power small cells compromising Wi-Fi through selective data offloading. A set of enhanced architectures and mechanisms are proposed through which traffic is offloaded selectively, based on the network conditions and QoS parameters required by the data traffic to improve the network performance. These algorithms use network information such as, available resources, channel conditions, service/application requirements, including delay and throughput, to make the offloading algorithms highly adaptive to these parameters. Novel algorithms are proposed at the core network level to dynamically monitor the network parameters and

wireless resource allocations in order to perform seamless offloading and maintain the service continuity. A new scheduling discipline is introduced in the MAC scheduler, i.e., the QoS-aware scheduling, where the average waiting time in the eNodeB buffer is computed for each type of traffic to satisfy the delay requirement of each application/service. Therefore, by using the service requirement as well as knowledge of both cellular and Wi-Fi network conditions, such as access network utilization, data traffic can be offloaded dynamically. The power and subcarrier allocation is also considered by the medium access control protocol that uses Orthogonal Frequency Division Multiple Access (OFDMA) resource allocation in the LTE scheduler. In that respect, a novel joint-resource allocation algorithm for LTE and Wi-Fi access networks is proposed which, together with IP flow mobility, performs seamless QoS-aware offloading. Furthermore, to make management plane functionalities more flexible and programmable, it is argued that the mobile communication architecture can be improved by applying the principles of Software-Defined Networking (SDN) and providing logically centralized control of the overall infrastructure, thus enabling programmability. The performance of the proposed schemes is investigated thoroughly. Various figures of merit, such as end-to-end throughput, end-to-end delay, offloading efficiencies and blocking ratio, are used to present the efficiency of the proposed algorithms in the wireless heterogeneous networks. The results show a significant overall improvement in the performance of the system through offloading to Wi-Fi.

1.2 Thesis Contributions

The contribution of this thesis will lead to the design of a smart offloading mechanism that is scalable and QoS-aware. The design issues of this mechanism are discussed in three main chapters that concern the concept of offloading. The first, will consider QoS of application/services in the core network; the second, will adapt the offloading policies for both user/service and network, while in the last chapter the policies which dynamically address scalability issues are derived and applied. The proposals that are elaborated in chapters 3-5 of the thesis are as follows:

1. The first proposed algorithm is presented in Chapter 3. Initially, the chapter highlights the impact of QoS in LTE resource allocation. A LTE/Wi-Fi seamless offload is proposed which enables seamless handover of data traffic between LTE and Wi-Fi, as well as selective IP traffic offloading, such as VoIP and video while supporting simultaneous LTE and Wi-Fi access. In the proposed model, LTE network resources, channel condition and user required services/applications are

monitored dynamically. A resource allocation framework is proposed to maximize throughput and minimize the waiting time in the queue among different classes of traffic by adapting the eNodeB scheduling algorithm to the required QoS of traffic type considered as waiting time in the buffer, and offloading the traffic when necessary.

2. A policy based offloading framework for cellular networks which is based on a cost function approach, is presented in Chapter 4. User centric, network centric and hybrid policies are discussed where the decision making is shared between the user and the network. Also, a novel mechanism for decision sharing between user and network, based on principles of autonomic networking, is proposed where policies are chosen dynamically according to the variation of network conditions and the operator strategies. Detailed simulation studies are conducted to validate the effectiveness of these policies in real networks.

3. An offloading mechanism which works through applying the abstraction of software defined networking (SDN) in the mobile backhaul, is proposed in Chapter 5 to provide programmable offloading policy derivation. The proposed mechanisms consider the real-time network conditions to derive the offloading policies and efficiently accommodate the traffic in both LTE and Wi-Fi network. Numerical results prove that the proposed approach can significantly improve the dropping rate of the incoming traffic by using more real-time and dynamic decisions for offloading.

1.3 Thesis Outline

The remainder of this thesis is organized as follows: Chapter 2 provides a literature review of the works related to the proposed research topic. The aim is to provide the technical background necessary to understand the problem area that is addressed in this thesis. Based on LTE architecture and offloading objectives, the chapter discuss the offloading techniques and different aspects of offloading. While all these categories are discussed briefly, the more detailed discussion is of policies, QoS-Aware and enhanced architectures in future networks, which is the main focus of this thesis. After reviewing the related literature in Chapter 2, the main contributions of the thesis are discussed in Chapters 3, 4, 5 and 6. Chapter 3 focuses on QoS-Aware resource allocation in offloading. Chapter 4 proposes a hybrid approach to offloading decision making. An enhanced architecture, based on the Software-Defined Networking approach to dynamic

policies, is discussed in Chapter 5. The concluding remarks, together with some avenues of future research, are detailed in Chapter 6.

The publications that are related to the contributions of the thesis are as follows.

Journals

1. **M. Amani**, T. Mahmoodi, and A. H. Aghvami, "Quality of Service Aware Mobile Data Offloading," IEEE Transactions on Communications, pp. 1-12. (Submitted).

2. **A. Aijaz**, M. Amani, and A. H. Aghvami, "Survey on Mobile Data Offloading," Communication Magazine, IEEE, vol. 16, no. 5, pp. 670-673, April, 2013.

Conference Proceedings

3. **M. Amani**, T.Mahmoodi, and A. H. Aghvami, "Programmable Policies for Mobile Data Offloading," IEEE ICC 2014(Submitted).

4. **M. Amani**, A. Aijaz, and A. H. Aghvami, "On Mobile Data Offloading Policies in Heterogenous Wireless Networks," IEEE VTC-Spring Jun. 2013 .

Chapter 2

Background Study

The unprecedented increase in data traffic in the last few years, is creating challenges for existing cellular networks. Mobile data offloading or simply data offloading has been extensively discussed as a key solution to alleviate congestion and make better use of available network resources. Most of the mobile operators have introduced and started to implement a mobile data offloading strategy. So far, Wi-Fi have emerged as the preferred offloading technologies due to the availability of Wi-Fi interface on smartphones and tablets. The goal is to dynamically select and offload the traffic toward a low-cost access network to alleviate, data congestion while maintaining Quality-of-Service (QoS) for customers and reducing the cost and impact of carrying capacity-hungry services on the mobile network. 3GPP and other body of standardization very actively have been developing a new standard to support the simultaneous use of cellular access networks such as LTE, femtocells, and non- 3GPP access such as Wi-Fi [2–4]. In this chapter, a full overview of wireless network evolution and different techniques to address the challenges of today's mobile networks are given. Section 2.1 briefly discussed the arising challenges of today's mobile networks. In Section 2.2, the trend of wireless technologies is elaborated, thus emergence of new generations of wireless are presented. In Section 2.3, the state of the art offloading techniques are detailed. As the focus of this thesis is Wi-Fi offloading, it is covered in relatively more detail compared with other emerging technologies. After explaining the challenges of mobile data offloading in this section, the solutions offered by 3GPP are detailed in Section 2.4. Section 2.4 shed the light on the main topic of the thesis and explain the Quality of Service (QoS) and Policy in Mobile Data Networks aiming to enhance the performance of the LTE system.

2.1 Today's Mobile Network Challenges

Mobile Internet access is becoming more and more popular today and the ubiquitous cellular data networks attracts millions of users, which is in turn create pressure on the limited spectrum of these networks. Consequently users are suffering from poor quality of service due to network struggle to provide the required service. There are a number of factors contributed to the rise of traffic. Not only the popularity of high-end devices such as labtops, tablets and smartphones can generate much more traffic than a phone with basic features, but also increasing contact time between device and network due to the improvement in the mobile network connection speed and mobile devices' battery life contributes to the traffic rise. The increase in mobile video content which has higher bit rates than other mobile contents is another factor for the data traffic growth. Higher data rates provided by current mobile networks, large screen sizes and optimization of video for mobile devices improves the user's viewing experience and increase the mobile video content which has higher bit rates and generates more traffic than other types of traffic. Availability of mobile broadband services with comparable speed and price to the fixed broadband also contributes to the traffic rise on mobile networks. A variety of access technologies has been deployed over the last few years. While the 2G cellular systems evolved in to 3G systems like UMTS providing wide area coverage, Long Term Evolution (LTE) facilitate the usage of wider spectrum and complement 3G networks to increase the capacity and handle more traffic. However, limitation of radio link improvement and available spectrum to the operators make it imperative to seek other solutions.

Coexisting of different wireless technologies in the same area could try to overcome some of the most common issues of wireless access networks such as limited bandwidth, incomplete coverage, and time-varying channel conditions. Users equipped with multiple interfaces could benefit from the highest possible quality of service unaware of the employed access technology. Therefore, switching from one access network to another could optimize the performance and resource utilization while the service continuity should be guaranteed in a transparent way. On this note, mobile data offloading or data offloading is considered as a key solution both by Mobile operators and vendors as the data traffic on mobile networks continues to grow rapidly which cause congestion and degradation of user experience. Other factors which inspired the data offloading are cost reduction, improve user experience, and new business opportunities. Most of the mobile operators started to consider and implement mobile data offloading to small cells and and Wi-Fi. Although there are many wireless access technologies in-

terworking architectures available in the literature, quality of service guarantee and seamless roaming are two major drawbacks of these proposed architectures. One of the important challenges in any data offloading solution is to make a real time decision to offload different user/device and service /application selectively considering network conditions and environment. The performance improvement of heterogenous network consisting of high power macro-cells and low power small cells such as Wi-Fi through selective data offloading is a challenging topic and attracted many research studies which is also motivation for this piece of research.

2.2 Evolution of Wireless Access Networks

Despite the fact that the traditional wireless and mobile networks developed to target the voice, new generation of smartphones and android along with 3G/4G enabled laptops and netbooks introduced huge data growth to these networks over the past few years. According to forecasts, global mobile data traffic will grow 13-fold from 2012 to 2017; in particular, global mobile data traffic will grow 3 times faster than fixed IP traffic [1]. While mobile network operators will carry the bulk of Internet traffic in the future, they face significant challenges in addressing the needs of increased traffic demand. To address the exponential rise of data usage, mobile networks has been evolved over the past few years and new technologies has been emerged not only to provide higher data rates but also support Quality of Service (QoS) for real-time services e.g. multimedia.

The evolution of wireless communication over the last decades is considered among one of the biggest engineering success remarks. The first generation (1G) of mobile radio systems based on analog transmission for voice services was introduced in the early 1980s. In Europe, Total Access Communications System (TACS) was introduced with 8 kbps data rate and the Frequency Division Multiple Access (FDMA) as its multiplexing technique. Soon after, the first version of the most popular wireless access technology, Global System for Mobile Telecommunications(GSM) drafted by European Telecommunications Standards Institute (ETSI) replaced 1G and referred as second generation (2G). GSM uses the licensed spectrum of 900 and 1800 MHz as the most common frequency bands and it is based on Time Division Multiple Access (TDMA). GSM was designed to support voice initially and expanded over time to include limited data services through circuit switched data transport. Furthermore, it is evolved to 2.5G, General Packet Radio Service (GPRS), to include packet data transmission pro-

tocols. GPRS packet data rate transmission were later increased via Enhanced Data rates for GSM Evolution (EDGE) through employing EDGE-compatible transceiver units and upgrading BSS to support this technology. However these technologies could not accomplish the data rate and QoS required by the new applications.
Both the research and standardization body continues their effort, and GSM standard is succeeded by the third generation (3G) Universal Mobile Telecommunications System (UMTS) standard developed by 3GPP and further evolved to fourth generation (4G) of wireless networks such as Long Term Evolution (LTE) and LTE Advanced (LTE-A). However, limitation of cellular network has proved that deploying only macro-cells in wireless network is only effective to provide coverage and capacity for voice and low data rate services and it is hard to meet the requirement of high data rate services. Instead Heterogenous Network (HetNet) compromising of multiple types of radio access nodes is accepted widely as a promising technique to meet the increasing traffic demand in mobile wireless networks. However, these technologies improve the coverage and capacity of macrocells, the interference between these access technologies is severe. Additionally users mobility pattern from one place to another and bursty nature of mobile data application, result in fluctuating the traffic very much from time to time. Although, The average utilization rate of individual access network is very low, these processing resources cannot be shared with other resources. Thus all the nodes must be designed to handle the maximum traffic, no matter the size of the average traffic. To address this problem Cloud Radio Access Network (C-RAN) is invented by re-defining the implementation and deployment method of cellular networks. Cloud Computing is followed by Software Defined Networking to address some of the scalability and management in today's telecommunication networks. In this section a brief introduction to the evolutionary trends of wireless networks including technologies such as WiMax, Wi-Fi and HetNets are reviewed .

2.2.1 UMTS

Universal Mobile Telecommunications System (UMTS) emerged as the most important third generation mobile cellular technology based on the GSM to provide data services by Third Generation Project Partnership (3GPP). UMTS employs Wide-band Code Division Multiple Access (WCDMA) radio access technology to offer greater spectral efficiency and bandwidth to mobile network operators where a pair of 5 MHz-wide channels typically is used for transmission in Frequency Division Duplex (FDD) mode. Spread-spectrum technology is employed where each transmitter is assigned a spreading code to allow multiple users to be multiplexed over the same physical channel. UMTS

specifies a complete network system, covering the radio access network UMTS Terrestrial Radio Access Network (UTRAN), the core network (Mobile Application Part, or MAP) and the authentication of users via Subscriber Identity Module (SIM) cards. 3G evolved later to High-speed Downlink Packet Access/High-speed Uplink Packet Access (HSDPA/HSUPA), therefore data rates could reach as high as 14.4 Mbps in the downlink direction and 5.76 Mbps in the uplink direction. The scheduling procedure was only performed by NodeB leading to faster resource management. The minimum Transmission Time Interval (TTI) was decreased from 10 ms to 2 ms in order to allow reduce latencies. Furthermore, Evolved HSPA (HSPA+) is expected to offer downlink data rates of 21 Mbps and uplink data rates of 11 Mbps.

2.2.2 LTE

Long Term Evolution (LTE) is the 3GPP specification for the fourth generation of mobile networks, also referred to Evolved UMTS Terrestrial Radio Access (E-UTRA). LTE represents a step forward for the wireless communications, targeting order-of-magnitude increases in the bit rates with respect to its predecessors by means of wider bandwidths and improved spectral efficiency. Beyond the improvement in bit rates, LTE aims to provide a highly efficient, low-latency and, packet-optimized radio access technology offering enhanced spectrum flexibility. The Orthogonal Frequency Division Multiplexing (OFDM) technique has been selected for the downlink and Single Carrier Frequency Division Multiple Access (SCFDMA) for the uplink. The downlink supports data modulation schemes QPSK, 16QAM, and 64QAM and the uplink supports BPSK, QPSK, 8PSK and 16QAM. At the network layer, a flatter architecture is being defined that represents the transition from the existing UTRA network which combines circuit and packet switching to an all-IP system.

2.2.3 LTE-Advanced

In 2008, 3GPP held two workshops, where the Requirements for Further Advancements for E-UTRA were discussed. The resulting technical report that has been published in [5], addresses the mobile systems whose capabilities go beyond those of IMT-2000, and called IMT-advanced. Some of the main objectives have been identified as interworking with other radio access systems and enhanced peak data rates to support advanced services and applications. It can provide data rate as high as 100 Mbps to 1Gbps for high and low mobility users consecutively.

2.2.4 HetNet

To overcome the limitation of traditional deployment of only macro-cells in cellular network and support higher data rate applications, HetNet is widely accepted to meet the requirement of the increasing traffic demand in mobile networks. HetNet compromises of multiple types of radio access nodes in a 3GPP LTE network such as macro Evolved Node B (eNodeB), pico eNodeB, femto eNodeB, and relay. In HetNet, macro eNodeBs provide coverage for low speed services, while small cell eNodeBs deployed in the macro-cell ensures hotspot coverage and capacity enhancement. The small coverage area of small cells means that lower number of users sharing the same cell comparing to macro-cell which gives more freedom to users to use the bandwidth with lower transmission power [6, 7].

2.2.5 Wireless Local Area Network (WLAN) and Wi-Fi

A wireless local area network (WLAN) links two or more devices typically using spread-spectrum or OFDM radio and usually provides connection to the wider internet through an access point. Moving users within a local coverage area of the WLAN still benefits from network connection. Most modern WLANs are marketed under the Wi-Fi , stands for Wireless Fidelity, which is a wireless connectivity solution based on IEEE 802.11 standards. It is primarily used for broadband access in indoor environments. Compared with conventional mobile communication technologies (e.g., UMTS, HSPA, LTE etc.), Wi-Fi provides higher data rates with short range wireless networking using the unlicensed 2.4 or 5.3 GHz band but with limited coverage and mobility. Nowadays, Wi-Fi is undergoing a paradigm shift towards ubiquity and outdoor/city-wide Wi-Fi networks are gaining popularity. WLAN specifications has been enhanced over the past few years which result in e.g. IEEE 802.11a that uses Orthogonal Frequency Division Multiplexing (OFDM) technique, IEEE 802.11n that supports Multiple Input, Multiple Output (MIMO), and IEEE 802.11s for mesh networking.

2.2.6 Worldwide inter-operability for Microwave Access (WiMAX)

WiMAX provides point-to-multipoint wireless communications based on the IEEE 802.16 standards [8] for wireless broadband access over large geographical regions. The original standard defined the radio frequency band in the range 10 to 60 GHz, and afterwards IEEE 802.16a [9] updated and added the specifications for 2 to 11 GHz range. Further enhancements have been added to the standards which introduced the use of Frequency Division Multiple Access (OFDMA) as the access method, and support of

MIMO in IEEE 802.16e [10]. This brings potential benefits in terms of coverage, self installation, power consumption, frequency re-use and bandwidth efficiency. WiMAX is the most energy-efficient pre-4G technique among LTE and HSPA+ .

2.2.7 Software Defined Networking

Due to the static nature, today's network cannot dynamically adapt to changing of traffic, application, and user demands. To apply any changes , the multiple switches, routers and protocol-based mechanisms should be modified. Thus it makes very difficult to apply a consistent policies such as access, security and QoS to the growing number of mobile users. On the other note, to serve users with different applications and performance needs, networks require additional devices to introduce new services which will result in higher expenses. Additionally, the huge data growth means increasing the capacity which in turns introduce the constant demand of scaling the network. The solution to all these challenges is software-defined wireless networking, whereby low cost wireless hardware such as small cells and Wi-Fi access points are extensively deployed, with cloud-based software dynamically optimizing the overall network performance and providing intelligent seamless mobility in the heterogenous networks consisting of HetNets and Wi-Fi access networks.

2.2.8 SDN for Cellular Data Networks

Software Defined networking is a network architecture where network control is decoupled from forwarding plane and is entirely programmable. Moving the control function which was tightly coupled to individual devices to accessible computing devices, allows the underlying infrastructure to be abstracted from applications and services. The control function is centralized in SDN controllers which they have a global view of the network. SDN enables the operators to control the entire network from a single logical point independent of the vendors which simplify the network design and operation. Additionally, network devices do not need to process various standards, they just perform what is instructed by SDN controllers. Centralizing the control function enables the network operators to configure, manage , secure, and optimize network devices through dynamic software programming. Thus it makes real-time modification and new services deployment more feasible in short period of time. SDN also supports a set of API which makes it easier to define and implement consistent policies across the network to meet the operators objectives. The Open Networking Foundation is introducing open APIs which promote multi-vendor management and facilitates virtual networking and

cloud services. Also open APIs between SDN and applications optimize the computing and resource management by abstracting the network from applications. some of the benefits of SDN are as follows:

- Centralized management of network consisting of multi-vendor devices
- Improved management by using common APIs to separate network details from applications
- Rapid innovation due to deliver new services without the need to configure individual devices
- Programming by operators, vendors and users using common programming environments
- Increased network reliability due to centralized management of network devices
- Better user experience as applications use the centralized network information to adapt the network behaviors to user needs
- Fine grained network control due to ability of applying wide range of policies at the session, user, device and application level

Furthermore, Open Flow based SDN address today's network challenges such as high bandwidth and dynamic nature of the applications and reduce the management and operation complexity. Some of the Open Flow

- Centralized Control of Multi-vendor Environment
- Reduced Complexity
- High Rate of Innovation
- Increased network reliability and security
- More granular network control
- better user experience

2.2.9 Cloud Computing

Cloud computing is a model for enabling convenient and ondemand network access to a shared pool of configurable computing resources. It has many characteristics, such as highly dynamic utility computing or elasticity, and the ability to access and pay for more or fewer resources, such as central processing units (CPUs) and storage, as demand increases or decreases. Virtualization technology and broadband connectivity are what makes this possible. In simple terms, virtualization allows traditionally course-grained physical resources to be split into fine-grained virtual resources that can be allocated on demand over a broadband connection.

2.3 Offloading & Offloading Techniques

Mobile data offloading or simply data offloading refers to the use of complementary network technologies and innovative techniques for delivery of data originally targeted for mobile/cellular networks in order to alleviate congestion and make better use of available network resources. The objective is to maintain Quality-of-Service (QoS) for customers, while also reducing the cost and impact of carrying capacity-hungry services on the mobile network. Mobile data offloading already becomes a key industry segment as the data traffic on mobile networks continues to increase rapidly. So far, Wi-Fi and femtocells have emerged as the preferred offloading technologies which are elaborated in this section along with other available solution.

2.3.1 Mobile Data Offloading via Wi-Fi

Wi-Fi comes as a natural solution for offloading due to built-in Wi-Fi capabilities of smartphones. Due to degradation of cellular services in overloaded areas, an increasing number of users are already using Wi-Fi to access Internet services for better experience. From service providers perspective, Wi-Fi is attractive because it allows data traffic to be shifted from expensive licensed bands to free unlicensed bands (2.4GHz and 5GHz). Studies have shown that expanding network using Wi-Fi is significantly less expensive compared to a network rollout.

There are three main approaches for operators to offload data traffic onto Wi-Fi networks, depending upon the level of integration between Wi-Fi and cellular networks. The first approach is the network bypass or the unmanaged data offloading in which case the users data is transparently moved onto the Wi-Fi network, whenever they are

in Wi-Fi coverage, completely bypassing the (cellular) core network for data services. Voice services on the other hand continue to be delivered via the core network. Whilst this approach seems attractive as it does not require the deployment of any network equipment, it has some drawbacks. Firstly, the operator loses visibility (and hence control) of its subscribers whenever they are on the Wi-Fi network. Secondly, the operator is unable to deliver any subscribed content (Blackberry, corporate VPN, ringtones etc.) leading to potential loss of revenue. Despite its disadvantages, this approach can be adopted as an immediate offloading solution due to its ease of deployment. It is also attractive from users perspective due to control over data connectivity. It should be noted that this approach is similar to the users switching on the Wi-Fi interface whenever they are in Wi-Fi coverage for better experience. The operator can deploy such an offloading solution by simply placing an application in handsets that switches on the Wi-Fi interface on detecting Wi-Fi coverage.

A managed data offloading approach can be adopted by those operators who do not want to lose control of their subscribers. This is achieved by placing an intelligent session aware gateway through which the subscribers Wi-Fi session traverses on its way to the Internet. Complete integration of cellular and Wi-Fi networks is not required in this case. While the operator gains control of subscribers, it still cannot deliver any subscribed content. On the other hand, an integrated data offloading approach provides the operator with full control over subscribers as well as the ability to deliver any subscribed content while the users are on the Wi-Fi network. This is achieved by the integration of cellular and Wi-Fi networks so that a bridge can be formed between the two networks through which data flow can be established.

There are two architectures for coupling cellular and Wi-Fi networks: loose coupling and tight coupling. In loose coupling architecture, the networks are independent requiring no major cooperation between them. The Wi-Fi network is connected indirectly to the cellular core network through an external IP network such as the Internet and service connectivity is provided by roaming between the two networks. On the other hand, in a tightly coupled system, the networks share a common core and majority of network functions such as vertical handover, resource management, and billing are controlled and managed centrally.

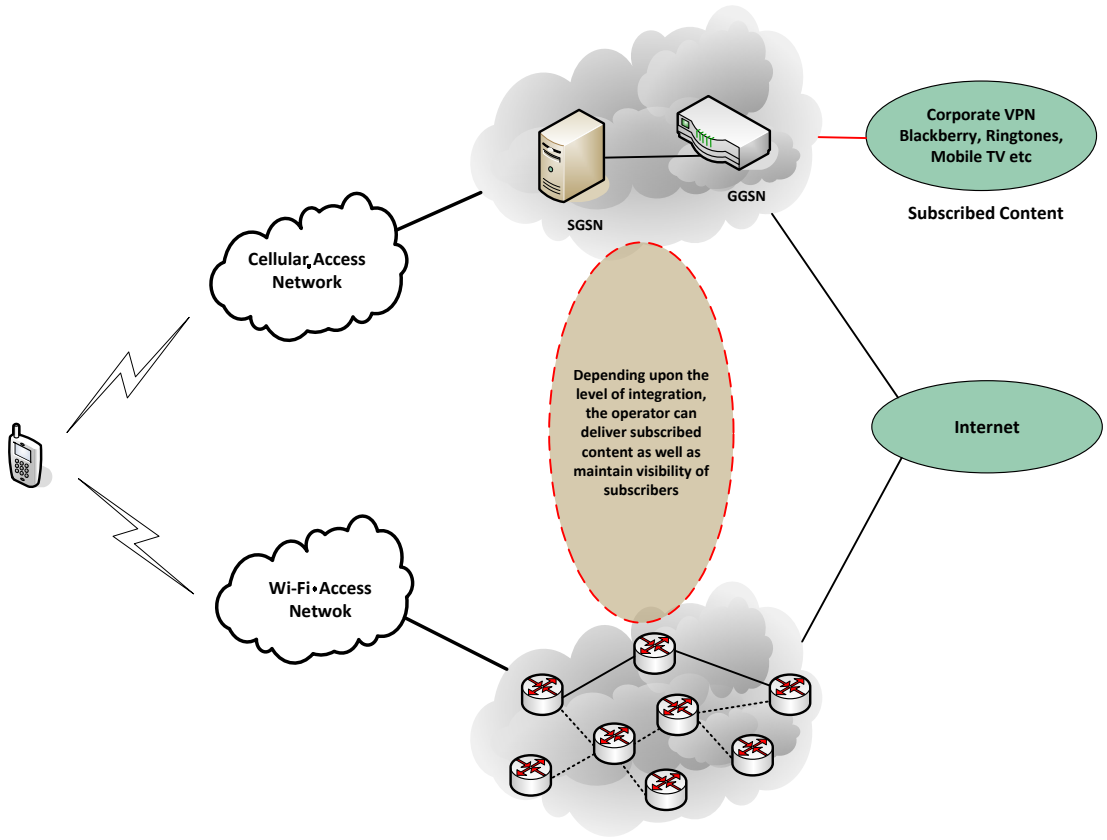

Figure 2.1: Mobile Data offloading via Wi-Fi

2.3.2 Mobile Data Offloading via Femtocells

A femtocell is a small, low-power cellular base station, typically designed for indoor use such as home or office.The more common term which is used in industry is small cell which includes femtocell as a subset. It connects to the service providers network via broadband (such as DSL) and allows the service provider to extend service coverage indoors especially in areas where access would otherwise be limited or unavailable. Femtocells are attractive to operators as they provide improvement in both coverage and capacity, especially indoors. The concept of femtocells is applicable to all standards including GSM, WCDMA, WiMAX, and LTE. Femtocells provide a highly effective method of easing the traffic carried by macro cellular network. The freed capacity improves the experience of customers on macro network whereas at the same time, users connected via femtocells experience improved performance due to usually better radio conditions.

Data offloading through femtocells is effective for a number of reasons, some of which are as follows [11]. Firstly the usage occurs primarily indoors (homes or offices). According to one of the published studies, 55% of data usage occurs in the

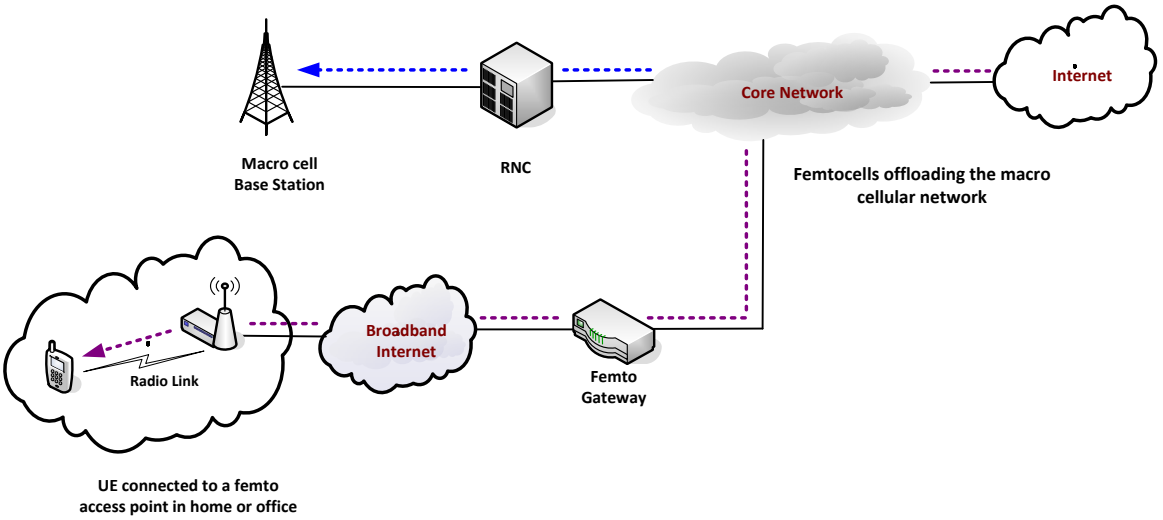

Figure 2.2: Mobile Data offloading via Femtocell

home and 26% occurs in the office. Thus the operators get the opportunity to offload heavy users through femtocells. Secondly, femtocells represent an operator deployed and managed service, and therefore provide a seamless experience to users. Thirdly, femtocells can be deployed quickly unlike traditional macro cellular deployments which take much longer due to site acquisition, purchase of radio infrastructure, backhaul, and other similar considerations. In femtocell environments, the traffic flows over the air interface to the femtocell (which is connected to the users broadband connection), then over the Internet to the operators core network and/or to other Internet destinations. Whenever a subscriber comes into the coverage of femtocell, the UE (user equipment) automatically associates with it. Traffic, which was previously flowing between the macro cell and the UE now flows through the femtocell and subscribers broadband connection. The femtocell not only offloads the Node B but also the RNC (Radio Network Controller) which further reduces the load on the macro cellular network. A new standard, currently under development, known as SIPTO (Selected IP Traffic Offload) [12] enables the operator to offload certain types of traffic at a network node close to the UEs location. The current standardization process mainly considers two types of policies for offloading: APN (Access Point Name) based and DPI (Deep Packet Inspection) based. However, it is important to mention that by implementing SIPTO, operators can offload the core network by allowing the traffic to flow directly from the femtocell to the Internet.

2.3.3 Advantages and Disadvantages of Wi-Fi and Femtocell

Both femtocells and Wi-Fi are major offloading solutions. As Wi-Fi operates in unlicensed bands, therefore operators have access to much larger free spectrum to cater for any size of Wi-Fi deployment. Femtocells, on the other hand require careful planning as they operate in costly (licensed) and limited spectrum bands. Femtocells capture 100% of traffic, whether it is voice or data and whether it originates from a feature phone, smart phone, or a laptop. This is usually not possible in case of Wi-Fi. Femtocells do not increase the power consumption on terminal side, whereas Wi-Fi enabled devices may experience increased battery drainage because of the power required to operate two radio interfaces. When it comes to data rates, Wi-Fi is the only technology that can deliver rates as high as 600 Mbps. Typically users on cellular networks need a lot of patience to download heavy multimedia files. Last but not the least, femtocells can provide guaranteed QoS using licensed band whereas Wi-Fi cannot.

2.3.4 Mobile Data offloading via Wimax

From data offloading perspective, WiMAX plays an indirect role by providing the backhaul for large scale Wi-Fi networks. An alternative solution is the Wi-Fi mesh technology which has been used for municipal Wi-Fi networks. However, point-to multipoint links using WiMAX for backhaul and Wi-Fi for access is a preferred solution as it does not introduce latency, reliability, and performance issues associated with the Wi-Fi mesh technology due to its non-direct, multihop nature. The authors in [13] describe a similar 3W data offloading strategy for offloading and efficiently handling the mobile data traffic. It comprises of WCDMA, Wi-Fi and WIBRO (Korean name for WiMAX) networks. Users can access the Internet with WCDMA packet service or Wi-Fi. The Wi-Fi network has two backhaul networks; conventional wired network and WIBRO network.

2.3.5 Mobile data Offloading via Opportunistic Communication

A recent approach to offload mobile data traffic using opportunistic communications has been proposed in [14–16]. Most of the information delivered over mobile networks comes from content service providers and may include multimedia newspapers, small computer games, weather reports etc. The service providers can benefit from the delay tolerant nature of such applications and may deliver the information to only a small

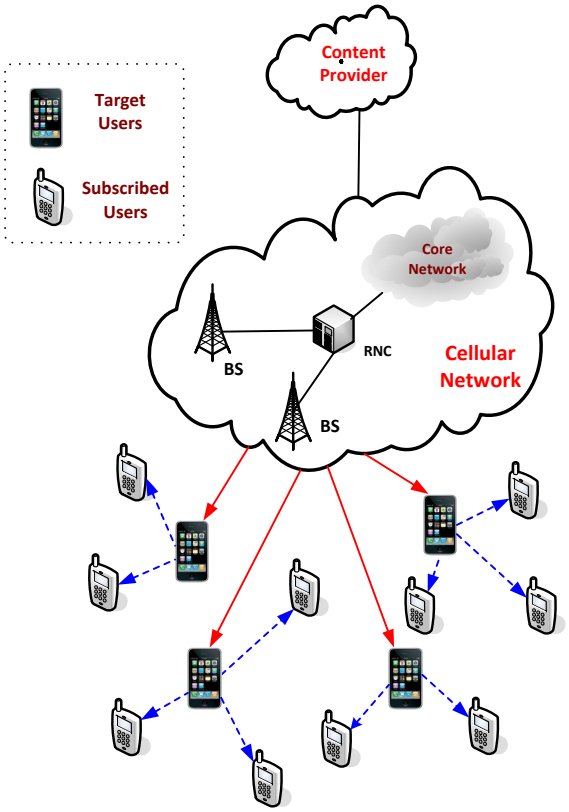

Figure 2.3: Mobile Data offloading via Opportunistic Communication

group of users (target users). The target users can further disseminate the information to subscribed users when their mobile phones are in proximity and can communicate opportunistically using Wi-Fi or Bluetooth technology. Apart from these two technologies, Device-to-Device (D2D) communication using cellular resources can also be employed for such opportunistic communication. Such an offloading approach is attractive as there is little or no monetary cost associated with it. However, it is challenging due to a number of reasons such as the heterogeneity of data traffic from service providers (varying in delay and content size), varied user demands and preferences for data traffic, incentives for target users and battery and storage constraints of mobile devices.

2.3.6 Mobile Data Offloading via IP Flow Mobility

IP flow mobility [17] is a recent technology that is currently being standardized in the Internet Engineering Task Force (IETF). This technology allows an operator to shift a single IP flow to a different radio access without disrupting any ongoing communication.

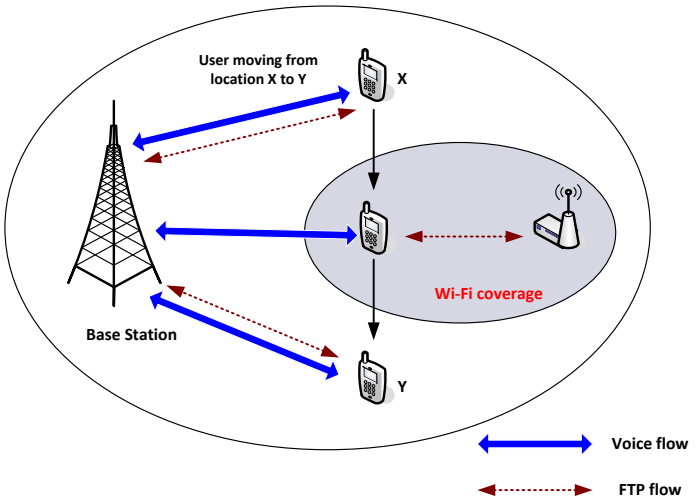

Figure 2.4: Mobile Data Offloading Via IP Flow Mobility

Consider a user connected to a cellular base station having multiple simultaneous flows (e.g., a voice call and a file download) moves into the coverage of a Wi-Fi access point (hotspot). The terminal or the network upon detection of the Wi-Fi access decides to shift the file download on the Wi-Fi network. Once the user leaves the Wi-Fi coverage, the file download is seamlessly shifted back to the cellular network. IP flow mobility provides the operators with a better traffic management solution by selectively offloading heavy users and thus alleviating congestion on their networks, while at the same time users enjoy high bandwidth connections and better experience.

2.3.7 Core network Offloading

In core network offloading approach, a gateway is placed in the data path between the RNC and the SGSN to offload the traffic destined for the Internet. The gateway detects, by using different identifiers such as Access Point Name (APN), if the incoming session request is for Internet bound services or operator bound services. In the former case, the gateway alters the traffic path, bypassing Packet Switched (PS) Core Network and forwards it directly to the nearest Internet peering point.

2.3.8 Media Optimization Solutions

Media optimization solutions, although not directly related to mobile data offloading, attempt to relieve network congestion by reducing the data traffic volume through a variety of techniques, and consequently improving user experience and network resource usage among different users and traffic types to ensure an overall optimal response to

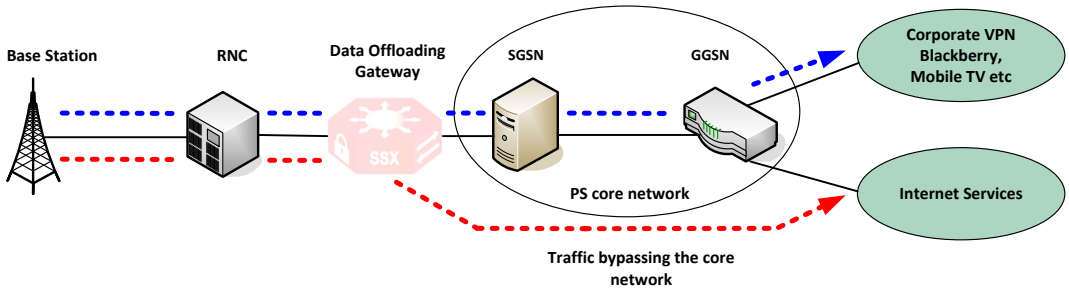

Figure 2.5: Mobile Data Offloading Via Core Network

the user. Such solutions are normally implemented by placing a media optimization server between the packet switched (PS) core network and the Internet. Some of the most commonly used media optimization solutions [18] are described as follows. []

Smart Video Caching

A popular video on the Internet can go viral in a short period of time. This increases the load on the mobile network infrastructure. One way of dealing with viral videos is to cache them closer to the user, instead of repeatedly retrieving from the server. To implement a more effective solution, the media optimization server should prefect the set of most popular videos, encode and store them according to device type and thus deliver the optimized videos from cache whenever they are requested.

Dynamic Bandwidth Shaping

The media optimization server dynamically adjusts the bit rate of encoded video according to the bandwidth available to the users. This enhances the end user experience by avoiding the screen freeze which happens when the incoming video bit rate is higher than available network bandwidth.

Web Optimization

Web optimization solutions provide faster browsing and more immediate access to the content by increasing data transfer rate over mobile networks, through different techniques such as analyzing traffic and storing the optimized version of most requested sites in the cache for quicker access, improving the load time for web pages by pipelining of HTTP requests, removing redundant white spaces and comments from HTML documents etc.

2.3.9 Challenges of Mobile Data Offloading

Some of the key challenges arise while implementing a data offloading solution, especially offloading through Wi-Fi. Such challenges must be addressed properly. The foremost challenge associated with data offloading is user experience. Service providers must ensure consistent user experience and service continuity, independent of the underlying offloading solution. This includes providing a transparent login across different networks to avoid any disruptions in service continuity. Since subscribers authentication data is present in 3GPP networks Home Subscriber Server (HSS) or HLR which cannot be accessed easily through non-3GPP networks such as Wi-Fi. An integrated offloading approach and SIM based authentication procedure by including an Authentication, Authorization and Accounting (AAA) server in the core network will ensure transparent sign on and seamless in session handover between 3GPP and Wi-Fi networks. The latter requires network readiness prior to device readiness.

Furthermore, as the Wi-Fi network cannot provide QoS guarantees, a QoS driven vertical handover from Wi-Fi to 3GPP networks is essential in order to ensure the QoS for users, especially when the Wi-Fi network is experiencing congestion. From users perspective, it is not suggested to keep both Wi-Fi and cellular interfaces simultaneously turned on due to significant battery drain, especially when the Wi-Fi interface is in idle mode. Moreover, the network cannot force a device to switch on the Wi-Fi. It is important to mention here that currently no outdoor Wi-Fi planning tools are available in the market. This creates challenges in optimal deployment of outdoor Wi-Fi access points. Also, roaming agreements between different Wi-Fi networks is an important issue that needs to be addressed if Wi-Fi offloading solutions are deployed on a large-scale.

A major challenge in femtocells is the interference management. The femtocell deployments create a two-tier network as a result of which, interference can be either co-tier or cross-tier. In case of co-tier interference, a femtocell causes interference to a neighboring femtocell which may be severe in case of dense deployments. The cross-tier interference results when a femtocell causes interference to the downlink of a nearby macro-cell user. Similarly a macro-cell user can cause interference on the uplink of a nearby femtocell. The interference problem becomes severe if the femtocell is operating in closed access mode.

Perhaps, the most important challenge in mobile data offloading through any of-

floading solution is to distinguish between different user segments and network conditions. Application, device, and subscriber awareness is required for making real time decisions regarding selective offloading and thus effectively managing the overall process.

2.4 Quality of Service (QoS) in Mobile Networks

Mobile broadband networks carry different type of services that share radio access and core network resources. Wireless networks must support mix of all different types of services such as real-time, non-real time and best effort services. Each service requires different QoS in terms of tolerable packet delay, minimum bit rates and acceptable packet loss rate. As mobile networks evolve to high-speed IP-based infrastructure, the wireless industry considers providing high quality services in addition to adding network capacity by developing QoS management framework. The designed techniques aim to manage network congestion, provide QoS for the applications and offer differentiated services to the users; Therefore 3GPP has defined a policy management to implement QoS in mobile networks [19, 20]. []

2.4.1 Policy Management

Policy management in general is deriving and applying resource allocation and network usage rules by the operator. Policies can be static or dynamic. Dynamic policies set the rules based on the network and application awareness, and include policy enforcement processes. Policy enforcement ensures the detection of various traffic types and applies the QoS rules accordingly. A solid policy strategy not only maintains the network performance and satisfies user demand, but also saves the network capacity. Mobile networks with limited resources benefit from dynamic policy management combined with other solutions such as offloading, and can satisfy the user's demand. Policy management enables the differentiation of services and user types; therefore facilitates granular control of the QoS for different services. Dynamic policy can allocate the resources based on the defined rules, e.g. reserve capacity to support delay sensitive application. Furthermore, they can control the priority, packet loss and delay in order to treat the specific services in a different manner. Also policies can limit the data rate on the network to provide fairness among the users and services for example, peer-to-peer file sharing is a high bandwidth, non-real time service which can consume considerable network resources and affect the networks ability to support real-time services.

2.4.2 QoS and Policy Management in LTE and EPC Networks

To define a policy control framework, 3GPP objective is to standardize a QoS and policy mechanisms which can provide service/application and subscriber differentiation. To this end a bearer model is defined to implement QoS. Furthermore a logical transmission path between user equipment (UE) and the external Packet Data Network (PDN) with well-defined QoS is considered. In this section, the 3GPP standard QoS terms which are used throughout this thesis is reviewed.

bearer model

A bearer is a logical edge-to-edge transmission path with defined QoS between UE and Packet Data Network Gateway (PDN-GW) and represents the level of granularity for QoS control in the Evolved Packet System (EPS)/E-UTRAN. It is considered as a basic element to separate the traffic in order to provide different treatment for traffic with different QoS requirements. A set of QoS parameters are associated with each bearer to describe the properties of the transport channel such as data rates, packet delay, packet loss, bit error rate, and scheduling policy in the radio base station. Depending on type of service such as a real-time or best effort service, a bearer includes two or four QoS parameters which are summarized as follows:

- QoS Class Indicator (QCI)
- Allocation and Retention Priority (ARP)
- Guaranteed Bit Rate (GBR) (real-time services only)
- Maximum Bit Rate (MBR) (real-time services only)

This means, all packet flows mapped to the same EPS bearer level receive the same packet-forwarding treatment.

QoS Class Indicator (QCI)

The QCI specifies the treatment of IP packets received on a specific bearer. The functional nodes in LTE such as PDN-GW or eNodeB handle the packet forwarding of traffic traversing a bearer. Several node specific parameters such as link layer configuration, queue management and scheduling weights are affected by the QCI values. 3GPP has defined a series of standardized QCI types, which are summarized in Table 2.1 [21]. Operators will consider three basic service classes such as voice, best effort

Table 2.1: Standard LTE QCIs

QCI	Resource Type	Priority	Packet delay budget (ms)	Packet error loss rate	Example services
1	GBR	2	100	10^{-2}	Conversational voice
2	GBR	4	150	10^{-3}	Conversational video (live streaming)
3	GBR	5	300	10^{-6}	Non-conversational video (buffered streaming)
4	GBR	3	50	10^{-3}	Real-time gaming
5	Non-GBR	1	100	10^{-6}	IMS signaling
6	Non-GBR	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
7	Non-GBR	6	300	10^{-6}	Video (buffered streaming)
8	Non-GBR	8	300	10^{-6}	TCP-based(for example, WWW, e-mail), chat, FTP, p2p file sharing, progressive video and others
9	Non-GBR	9	300	10^{-6}	

data, and control signalling to for the first deployments, and gradually will add dedicated bearers to provide premium services such as high quality video streaming or conversational voice.

Allocation and Retention Priority (ARP)

ARP which is used in bearer establishment is an important decision variable and becomes very important when the network is congested and particularly in handover situations when the subscriber roams to a heavily congested cell. In this case ARP is used to drop or downgrade lower priority bearers. The network will observe the ARP to decide whether a new dedicated bearers can be established through the radio base station.

Guaranteed Bit Rate and non-GBR Bearers

The bearers are categorized to two major types as follows:

- Guaranteed Bit Rate (GBR): Real-time services such as voice and video use GBR bearers. A GBR bearer has a minimum bandwidth requirement which is reserved by the network, and always consumes resources in a radio base station regardless of whether it is used or not. GBR bearers should not experience packet loss on the radio link or the IP network due to congestion. Real-time services typically require low latency and jitter which is defined for GBR bearers.

- Non-Guaranteed Bit Rate(non-GBR): Best-effort services such as file downloads, email, and Internet browsing use non-GBR bearers. These bearers will experience packet loss when a network is congested. Non-GBR bearers do not have specific network bandwidth allocation. On the other note, instead of a maximum bit rate on a per-bearer basis for non-GBR bearers, an Aggregate Maximum Bit Rate (AMBR) on a per-subscriber basis will be specified for all non-GBR bearers.

Service Data Flows (SDF)

The IP packets belongs to user service type such as email, browsing, etc. are specified by SDFs which are bound to specific bearers based on policies defined by the network operator. User Equipment (UE) is using traffic flow templates (TFT) to bind services to bearers at the PDN-GW. TFTs contain packet filtering information to identify and map packets to specific bearers. The filters can be configured by the network operator but they will contain at least five following parameters referred to 5-tuple.

- The source IP address
- The destination IP address
- The source port number
- The destination port number
- The protocol identification (i.e., TCP or UDP).

The Policy and Charging Enforcement Function (PCEF) in the PDN-GW [20] filters packets coming from external networks such as Internet or VPNs using TFTs.
Each IP packet entering the system is provided with a tunnel header on the different system interfaces. This tunnel header contains the bearer identifier so that the network nodes can associate the packet with the correct QoS parameters. In the transport network, the tunnel header further contains a Diffserv Code Point (DSCP) value. Since the bearer is the basic enabler for traffic separation, it provides differential treatment for traffic with differing QoS requirements.

2.4.3 Policy Decision Definition

Policy decision is classifying and managing the service flows based on the QoS parameters or other policy control attributes. Currently 3GPP just defines the 5-tuple set

described in the previous section as a classifier parameters for flows. This classification can be extended to other parameters such as application, location and network condition. The key terms defined by 3GPP standard are as follows:

- **Policy Decisions** are Policy Control and Charging (PCC) rules and IP-Connectivity Access Network (IP-CAN) bearer attributes provided by the Policy and Charging Rules Function (PCRF) to the Policy and Charging Enforcement Function (PCEF) for policy and charging control purposes.
- **PCC Rules** includes a set of information which is used to detect a service data flow and provide parameters for policy control and/or charging control.
- **Policy Control** includes QoS and charging control indicated by PCRF to the PCEF to control the IP-CAN bearer.
- **Charging Control** is the procedure where the packets belonging to a service data flow are associated to a charging key and online or offline charging are applied accordingly.

Furthermore, 3GPP is adding Traffic Detection Function(TDF) to PCEF in data plane management entities to identify and detect traffic details such as video or audio being used as data [22].

2.4.4 Centralized Policy Control

In centralized decision making process, a central entity is managing the policy control and charging which is referred as Policy Decision Point or PDP [23]. The PDP for network access control is implemented on the server within the network where all the information such as network coverage, mobility, and current load are available. In mobile standards, PDP makes a decision based on the data plane conditions reported by PCEF/TDF and instructs PCEF/TDF what policy actions to be taken.

2.4.5 Distributed Policy Control

In distributed decision making process, the local entities in data plane such as PCEF/TDF are allowed to make policy decisions when applicable, otherwise the data plane conditions are passed to the central policy entity in the control plane to make a decision.In practice, the data-plane entities include both measurement and control functionality. In distributed architecture, data plane performs data traffic identification and management by applying various techniques such as bandwidth-limiting in addition to monitor

the real-time usage. In this model both PEP and PDP are performed in data plane. In addition, a central policy entity located in the control plane coordinates information and policy uniformly for various types of interfaces, technologies and vendors. A distinguished feature of the distributed architecture is that most of the decision are taken by PDP/PEP entities. In some cases which more information is needed, the decision is taken by control plane.

2.4.6 Functional Elements in Implementing Policy and QoS in EPC

Multiple nodes in the EPC and LTE access play a role in implementing QoS and policy management. The PCRF located in PDN-GW dynamically manages and controls data sessions and also provides an interface for billing and charging systems. Furthermore, PCRF facilitate non-3GPP networks such as Wi-Fi or fixed broadband to access 3GPP LTE network. PCRF is the LTE policy manager which takes the operator policies, network information and user's profile stored in the HSS to make policy decisions. The HSS is a central data base and stores all the subscriber-related information such as QoS profile, associated MME, mobility management functionality, user authentication, and access authorization. All the PCC rules are distributed to the PCEF in the PDN GW. The PCEF enforces policy decisions by establishing bearers, mapping service data flows to bearers, and performing traffic policing and shaping, the details of interactions between PCC nodes can be found in [24]. The PDN-GW maps bearers to the underlying transport network. The transport network will typically be Ethernet based, and may use MPLS. The transport is not aware of the bearer concept and will use standard IP QoS techniques, such as DiffServ. The eNodeB is the radio base station in LTE and it plays a critical role in end-to-end QoS and policy enforcement. The eNodeB performs uplink and downlink rate policing, as well as RF radio resource scheduling. It uses ARP when allocating bearer resources. The effectiveness of radio resource scheduling algorithms in eNodeBs has a tremendous impact on service quality and overall network performance. Basically, eNodeB products can be distinguished from other competitive products on this basis. Like PDN-GW, the eNodeB maps bearer traffic to the underlying IP transport network. The UE also plays a role in policy by performing initial mapping of service data flows to bearers in uplink.

Chapter 3

QoS-Aware Mobile Data Offloading

In wireless mobile networks, demand for a large bandwidth and a high quality of service has increased rapidly to support multi-play of services. Long Term Evolution,(LTE) has emerged as one of the most promising cellular networks to support the next generation of mobile broadband access networks. However, the exponential growth of mobile data services demand has motivated operators to explore alternative networks to cellular networks to handle the volume of traffic so as to provide an acceptable user experience. In addition, the next generation of mobile wireless networks have been conceived to provide an 'always best connected' (ABC) service to users who can be connected to multiple access networks at the same time [25]. In this way, mobile data offloading to Wi-Fi networks has been introduced to address the data growth challenges and to use end user multiple interfaces simultaneously. In this context, part of the service providers' planning is to decide which of the actual flows should be serviced via which of the wireless access points. This chapter proposes a resource allocation strategy which presents a QoS-aware scheduling algorithm in LTE eNodeB and uses the integration of Wi-Fi access networks into LTE core networks to increase the system throughput while maintaining the user/service required QoS.

There are only a few studies extant in the current literature on offloading. For example, in [26], the delay tolerance of applications is leveraged and switching is performed to augment 3G with Wi-Fi in a mobile environment. An experimental testbed and simulation studies are conducted in [27] in order to increase the offloading efficiency through delayed transmission. Furthermore, a study in [28] reported that many operators manage the congestion issues by performing prioritization of heavy users in wireless networks. Authors in [14] suggested delaying the delivery of information over cellular networks and offloading it through free, opportunistic communications. In

[17], the advantages and drawbacks of two approaches which enable service data flow mobility based on IETF and 3GPP standards, are presented. A combined cellular and an opportunistic network framework is proposed in [29] for spatial uplink mobile data offloading. This algorithm, spatially offload uplink traffic in a traffic concentration area to non-congested areas based on user behavior prediction. [30] proposed an analytical model to quantify amount of 3G resources saved through offloading and measure the deadline assurance for measuring the quality of user experience with PCC support. A qualitative survey of mobile datat offloading is presented in [31] and a cross-learning approach in introduced in [32]. [33] formulate the offloading problem as a reverse auction and solved it with greedy algorithm. All these studies prove that a significant amount of data traffic can be offloaded onto Wi-Fi networks, if properly deployed; however, the QoS required by the application/services to satisfy user experience has not been taken into consideration.
Before any sophisticated deployment on Wi-Fi offloading areas, it is crucial to implement an appropriate mechanism to control the QoS and raise the quality of the experience for the user.

LTE presents a significant change to the 3G UMTS/HSDPA radio access and core network by using orthogonal frequency division multiple access (OFDMA) for downlink transmission and supports higher data rate and lower latency [34] The former is supported by OFDMA while the latter is achieved by using a transmission time interval of 1 ms combined with features such as hybrid-ARQ in air interface. Furthermore, the base station, eNodeB, performs all the radio access network functions, such as packet scheduling, which is one of the most important LTE features in supporting the QoS. A MAC scheduler in the eNodeB MAC sublayer is responsible for scheduling transmissions over the LTE air interface in both downlink and uplink directions and physical resource allocation. The MAC Scheduler is a complex component and presents a number of design challenges. Therefore this chapter presents a QoS-aware packet scheduling algorithm, which considers the QoS requirement of delays for various traffic classes, channel conditions and fairness. This algorithm monitors the delay experienced by each traffic class in the eNodeB buffer and how closed they are, to allow maximum delay. This algorithm increases the number of transmitted flows while satisfying the required delay and furthermore, increases the system throughput by offloading the best effort, and the traffic which cannot be transmitted within the allowed maximum delay on a Wi-Fi network.

1. A combined QoS-aware downlink LTE resource allocation and offloading mechanism is proposed.
2. The proposed scheduler performs resource allocation taking into account throughput maximization, fairness, and the queuing delay to take care of mixture of optimization objectives.
3. In the scheduling algorithm, physical resource blocks (PRBs) are selected dynamically according to the results of the proposed utility function.
4. The QoS-aware scheduling scheme introduces a new scheduling discipline in the MAC scheduler to ensure the real-time traffic receiving their desired QoS.
5. On the other hand, maximum possible throughput can be achieved through offloading for best effort traffic which does not have any QoS requirement.
6. Through extensive simulation studies, it has been shown that this novel mechanism considerably improves the network throughput for mixed classes of traffic.

The remainder of this chapter is organized as follows: Section 3.1, provides an overview of the literature on the LTE, LTE MAC scheduler and seamless LTE/Wi-Fi offloading through IP flow mobility. Section 3.2 elaborates on the motivation for this chapter. In Section 3.3, the resource allocation and baseline assumptions for the QoS-aware scheduling and offloading problem are detailed. After full overview of the simulation parameters, the performance of the proposed algorithm is investigated in Section 3.4 and numerical results are presented respectively. Finally, Section 3.5 provides a conclusion.

3.1 Background Study

Before examining the proposed algorithm and system model, a brief overview of the resource allocation in LTE and LTE/Wi-Fi interworking is given. Then, the LTE MAC scheduler and scheduling of different classes of traffic in LTE are briefly reviewed. Some of the existing scheduling algorithms for LTE and the challenge to provide QoS are stated as well.

3.1.1 LTE Cellular Network

In LTE, base station (eNodeB) is connected to the Internet via IP networking equipment, such as Serving Gateway (S-GW), and Packet Data Network Gateway (PDN-

GW), which are acting as mobility anchors for the user plane during handovers. User Equipment (UE) connects to eNodeB which routes the traffic via S-GW using GPRS Tunneling Protocol (GTP).

The eNodeB has a critical role in end-to-end QoS and policy enforcement. It performs uplink and downlink rate policing, as well as RF radio resource scheduling. When congestion occurs, eNodeB reduces the subscribers' max rate according to their profile in coordination with PDN-GW. Service quality and overall network performance in LTE relies heavily on the radio resource scheduling algorithms in eNodeB. 3GPP has defined a bearer model to provide a logical, edge-to-edge transmission path with defined QoS between the user equipment (UE) and packet data network gateway (PDN-GW) [35]. Each bearer is associated with a set of QoS parameters that describe the properties of the transport channel, including bit rates, packet delay, packet loss, bit error rate, and scheduling policy in the eNodeB. Each bearer has a QoS Class Indicator (QCI), which specifies the treatment of IP packets received on a specific bearer. QCI values have an impact on link layer configuration, scheduling weights, and queue management. The 3GPP has defined a series of standardized QCI types, which are summarized in Table 2.1 in [21]. LTE uses Orthogonal Frequency Division Multiplexing (OFDM) for downlink to overcome multi-path fading problems in UMTS and, as well as multi-carrier transmission, transmits data over narrow band carriers instead of spreading one signal over the complete 5MHz carrier bandwidth. The physical resources in LTE are divided in both frequency and time domain. The minimum time unit for scheduling is Transmission Time Interval (TTI) with 1 ms duration (TTI) and consists of two time slots. The LTE frame is formed by 10 consecutive TTIs lasting 10 ms. The LTE downlink physical resource is shown in Figure 3.1 as a time-frequency grid.

In frequency domain, the whole bandwidth is divided into sub-channels; each sub-channel corresponds to 12 consecutive, equally spaced sub-carriers. A physical radio resource which spans 12 consecutive sub-carriers at a sub-carrier spacing of 15 kHz in frequency domain, and 7 OFDM consecutive symbols over slot duration of 0.5 ms in time domain for the short cyclic prefix, is called a Resource Block (RB).

Therefore, one RB has 84 (12 sub-carriers $\times$ 7 = 84) resource elements corresponding to one slot in the time domain and 180 kHz sub-channel (12 sub-carriers $\times$ 15 kHz spacing = 180) in the frequency domain. The RB size is the same for all bandwidths; however, the number of available physical RBs depends on the transmission bandwidth

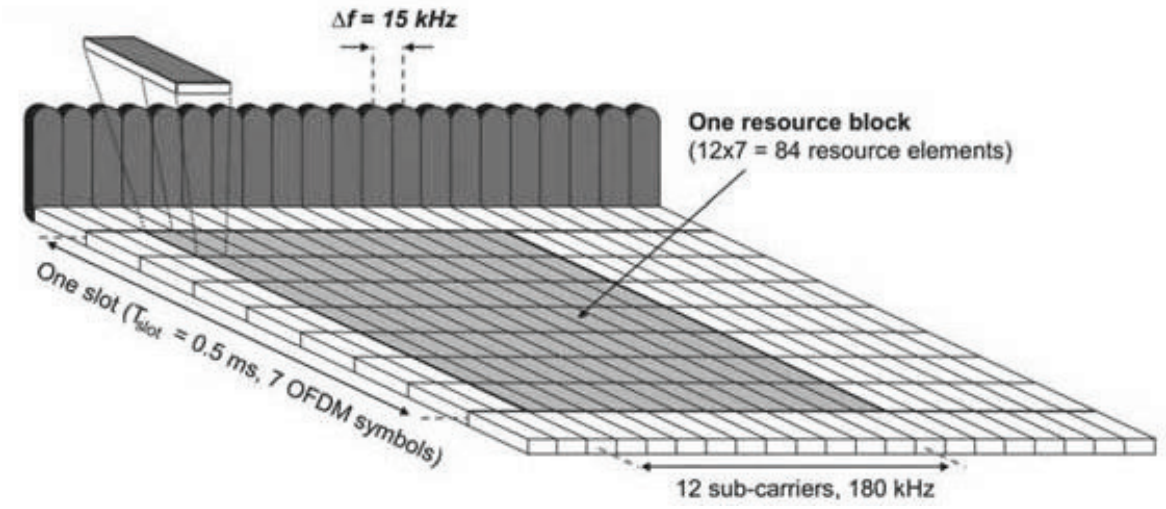

Figure 3.1: LTE Resource Block

and varies from 6 for bandwidth of 1.4 MHz, to 100 for a bandwidth of 20 MHz in frequency domain. The scheduler in eNodeB allocates the RBs to each user. The more RBs the user gets, and the higher modulation in each RB, points to the higher bit-rate. The allocation of RBs at any point to each user depends on advanced scheduling mechanisms in the frequency and time dimensions, which will be discussed in Section 3.1.6 and 3.4.

3.1.2 Seamless Wi-Fi Offloading

The wide availability of Wi-Fi at home, through various hotspots, and also in devices such as smartphones, laptops and netbooks that typically consume a large portion of resources, emphasizes that a mechanism which offloads data traffic from cellular network to Wi-Fi is very compelling for operators who want to optimize their resource usage. Mobile operators should be able to control which traffic is offloaded over Wi-Fi and which traffic is kept on a cellular network. For example, IP flows related to VoIP can be maintained on cellular networks to leverage QoS capabilities, while IP flows related to best effort (BE) internet traffic can be offloaded to Wi-Fi. First, a client-server based protocol, Dual Stack Mobile IP (DSMIP) is introduced by 3GPP in Release 8 [36] to enable seamless handover between cellular and Wi-Fi. IETF specifies DSMIP as a mobility protocol that provides IP address preservation for IPv4 and IPv6 sessions, allowing the users to roam in IP access independently [37]. The requirements to use this protocol are: 1) the cellular radio access network must support a Home Agent(HA);

and 2) DSMIP is available in both HA (i.e. server)and UE (i.e. client). Wi-Fi does not need to support DSMIP and HA functionality can be implemented in PDN-GW in LTE. The HA is the means to bind the node's permanent IP address, Home Address (HoA), to the local address which is based on its location, Care-of-Address (CoA). Therefore, the exposed IP address to the application remains the same. Seamless offloading in LTE using DSMIP can be implemented over S2c interface. DSMIP preserves the IP address whenever the network is changed, therefore, it provides a better user experience. To further improve the flexibility of DSMIP solutions, a 3GPP Release 10 introduces IP Flow Mobility that enables seamless movement of selected IP traffic while simultaneous access to cellular and Wi-Fi is supported.

3.1.3 IP Flow Mobility

A DSMIP solution in 3GPP Release 8 provides seamless Wi-Fi offloading where all the traffic is offloaded to Wi-Fi. However, it may be desirable that just some traffic is offloaded in some scenarios while other traffic is maintained over a cellular network. This needs an extension to DSMIP which will allow multiple addresses registration simultaneously. IP flow mobility allows selective movement of flows between different access networks of different technologies and handles IP flows separately within a PDN connection. IP Flow Mobility enables the dynamic allocation of different IP flows to different access networks based on their requirements, and allows the user to connect to two access networks (3G/4G and WiFi) simultaneously and forward/receive packets belonging to different flows through different access networks. Therefore, IP Flow Mobility is capable of registering multiple local addresses (i.e. CoAs) to a single permanent address (i.e. HoA) as is shown in Figure 3.2 Like all IP preservation mechanisms, IP Flow Mobility requires that traffic be routed through a central gateway (i.e. HA in PDN-GW); the traffic needs to pass through this gateway whenever mobility and session continuity are required. IP flow mobility facilitates the utilization of mobile devices which are equipped with multiple interfaces. An operator can control and configure the offload mechanism by considering the traffic characteristics, and optimize the resource utilization through load balancing among available resources. For example, the operator policy can indicate that all http traffic should offload to Wi-Fi while media should be handled by a cellular network.

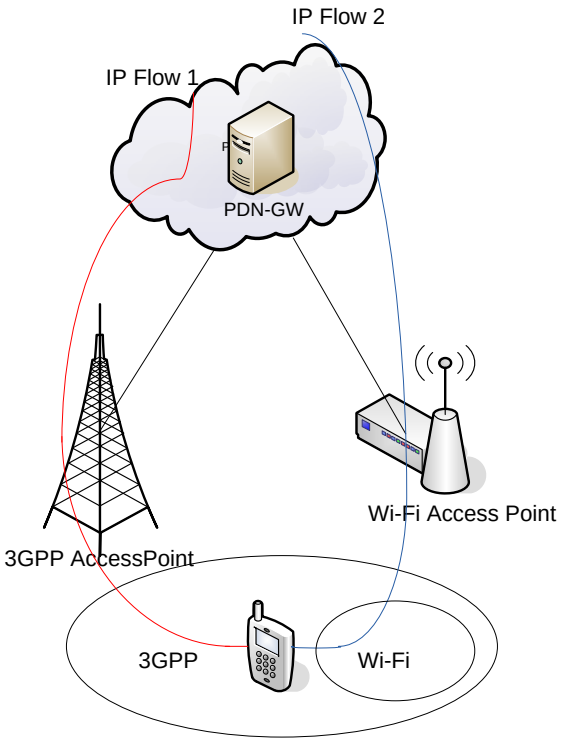

Figure 3.2: IP flow Mobility

3.1.4 Establish IP Flow with Service Continuity

The main objective of IP flow mobility is to establish IP flows belonging to an active PDN connection over multiple accesses and maintain service continuity. It is assumed that UE is under the coverage of both 3GPP and Wi-Fi access networks, therefore, simultaneous communication using multiple access networks is possible. An IP flow can move from the 3GPP access to the Wi-Fi access network towards the same PDN connection and vice versa. Either 3GPP or Wi-Fi access networks may/may not have any active IP sessions through it. Also an IP flow can be removed from the active PDN connections selectively, when UE has simultaneous active IP sessions via both access systems. In the case of loss of coverage, the UE moves all traffic associated with one access to another access and disconnects from one access. Each of the access systems can have one or more active IP sessions.

When the mobile terminal is under the coverage of both 3GPP and Wi-Fi access networks and both radio access interfaces are active simultaneously, IP flows can be transferred between access networks selectively. This means that the multi-radio interface mobile terminal can have IP flows towards the 3GPP access networks as well as IP flows towards Wi-Fi access networks. As it is shown in Figure 3.2, an IP flow can be moved from/to either of the access systems (from 3GPP to Wi-Fi or vice versa).

When UE goes out of the coverage of the one of the access network, all IP flows can be moved to the other access network to maintain the service continuity.

3.1.5 IP Flow Mobility Scenarios

In this section, IP flow mobility case studies are presented. These scenarios assume that UE is connected to the EPS via different access networks simultaneously, and IP flows can be sent through different access networks. Operator policies, the user preference, and the application characteristics, are some of the decision parameters in routing flows. For example, in Video Telephony calls, the conversational voice which is the heard real-time can be routed via 3GPP access, while the conversational video (live streaming) is the soft real-time and can routed through the non-3GPP access. The peer-to-peer download and media file synchronization are also routed through the non-3GPP access.

Case Study 1

In this scenario, it is assumed that UE only has 3GPP access on the way from office to home and is simultaneously accessing different services with different characteristics in terms of QoS requirements and bandwidth. For example, UE runs a web browsing session as well as a video telephony call consisting of conversational voice and non-conversational video streaming sessions. When UE reaches home, a device selects non-3GPP access i.e. Wi-Fi, and some of the current running services, such as web browsing sessions and non-conversational video streaming sessions, will be offloaded over a Wi-Fi access network based on the applications requirements, personal preferences, etc. Some of these flows may be from the same application but routed separately over 3GPP and non-3GPP access networks. Assume that, in the middle of the IP sessions, UE's device starts a non-real time FTP file synchronization with a backup server through Wi-Fi access network: with this huge amount of best effort traffic, the Wi-Fi access network becomes congested and therefore, the non-conversational video streaming session will be affected and does not get the required level of QoS, such as throughput and delay. This initiates the IP flow to move back to the 3GPP access. Later on, when the FTP file synchronization is done, the non-conversational video streaming session will be moved back to Wi-Fi. When UE loses the Wi-Fi connectivity, all the IP flows need to be moved to the 3GPP access, since it is the only access available. When UE comes to another area where both the 3GPP and Wi-Fi coverage are available, non-conversational video streaming, peer-to-peer download and

the non-real time FTP file synchronization are moved back to the Wi-Fi access.

Some of the advantages of offloading IP flows from the 3GPP access to the access are as follows:

1. Load balancing of the 3GPP access usage by using Wi-Fi access as a complementary access.
2. Optimal utilization of available radio resources
3. An increase throughout for IP flows with high throughput requirement.

Case Study 2

In this scenario, it is assumed that UE has a Skype session, i.e. a conversational voice session(VoIP) combined with a conversational video. During this session, UE also runs web browsing sessions (best effort) as well as a non-conversational video stream, such as YouTube. Based on the network operator policy, the VoIP flow and conversational video are routed via 3GPP access, while the non-conversational video and best effort IP flows are routed via Wi-Fi access network. Meanwhile, UE's device starts FTP file synchronization with a backup server (best effort) through Wi-Fi access and as a result WiFi access network becomes congested and the non-conversational video flows are moved back to the 3GPP access.

Later, HTTP server response time for the web browsing (best effort) is also increased and causes the best effort web browsing to move back to the 3GPP access network. Therefore, only the FTP file synchronization runs over a Wi-Fi access network. Finally, when the FTP file synchronization completes, the non-conversational video and web browsing are moved back to Wi-Fi access.

3.1.6 LTE MAC Scheduler

The LTE network is based on the Evolved Packet Core framework for delivering voice and data services. All services in LTE are provided as packet services, including voice services, which means real-time and non-real time services are multiplexed over the air interface and core network [38].

The benefits of LTE can only be recognized by networks which accommodate a mix of real-time and non-real time services dynamically; therefore, an end-to-end QoS is

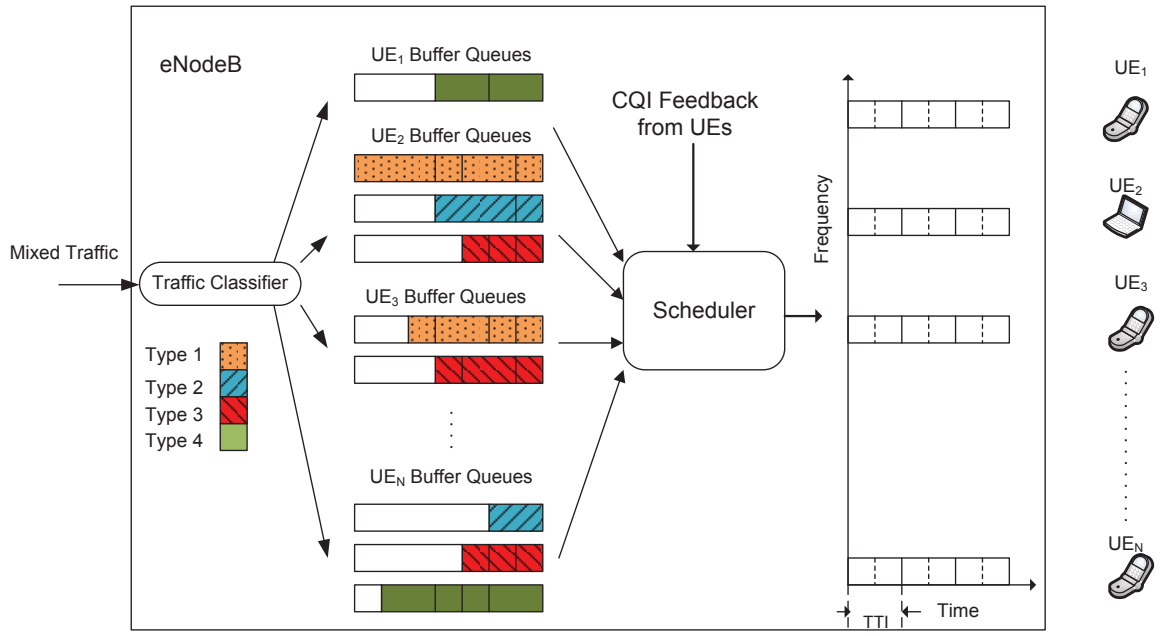

Figure 3.3: eNodeB Scheduler

defined in order to support such a mix of services. As a result of this, the air interface introduces an end-to-end QoS challenge which considers the time varying wireless channel conditions. User-plane and control-plane traffic travels over the air interface using bearers [39]. Bearers support various types of service with different QoS profiles, hence schedulers in MAC layer are used to guarantee QoS requirements for different types of applications. The MAC scheduler in the eNodeB MAC sublayer determines the uplink and downlink channel allocation of LTE air interface dynamically [40]. It schedules the packets based on QoS profile information, such as delay requirements, minimum bit rates and priority, which is received from the upper layers. In principle, the eNodeB allocates downlink or uplink available radio resources to each UE and its radio bearers based respectively on the downlink data buffered in the eNodeB, and on Buffer Status Reports (BSRs) received from the UE to achieve the negotiated QoS for each radio bearer. The eNodeB MAC scheduler is shown in Figure 3.3.

The incoming packets to eNodeB are classified into 9 classes according to QCI and stored in logical buffers for each user of a specific traffic class. The scheduler decides the head of line (HOL) packets transmission based on the scheduling policy in each TTI. In addition to the QCI of each packet, the scheduler considers the channel conditions of each user in scheduling. Therefore, UEs regularly measure channel conditions using reference symbols and report channel quality indicators (CQI) through uplink channels

to schedulers in eNodeB. The link adaptation module uses the adaptive modulation and coding (AMC) scheme to select the most suitable modulation scheme and coding rate (MCS) for each user, in order to maximize the spectral efficiency.

3.1.7 Related Works

The MAC scheduler design and algorithms used to allocate the radio resources has a critical impact on the eNodeB and overall LTE performance, therefore, finding an algorithm which will distribute the available resources to users attracts many researchers. These algorithms should consider the required QoS, channel conditions and data source behaviour. The problem is even more complex in the presence of various types of traffic with different requirements in terms of delay and bandwidth.
The existing resource allocation algorithms are categorized to real-time and non-real time and mixed traffic. For non-real time flows, the user experience is normally modelled through a concave thus increasing the utility function of the average rate experienced by a flow. The main focus of most of the previous works has been on [41] and [42] and is concerned with maximizing the sum rate, or minimum rate guarantee without considering any fairness among flows. Real-time traffic is typically modelled by a random packet arrival process to a queue, and the packets have a delay requirement to be met. The scheduling should ensure that the queue length does not grow without any boundaries via stabilizing policies. In all stabilizing policies categorized in [43–45], the average rate is equal to the mean arrival rate of the flow, while the delay distribution can be very different. However, to meet the flow delay and QoS requirement, it is not sufficient only to guarantee a minimum average rate for the flow. Approaches such as Maximum Throughput(MT), Proportional Fairness(PF), Weighted Round Robin [46–48] do not strictly consider the delay requirement, therefore, they cannot be applied in real-time traffic. Mixed traffic scheduling algorithms are also extensively studied in the literature. Some of the works such as [49, 50] propose a strict priority across classes of traffic which leads to a better performance with sub-optimal resource utilization. In [51] a scheduling policy gives equal priority to all packets until they are close to their deadline, then, such a packet benefits from a higher scheduling priority. In this policy, channel conditions, queue states and delay deadlines are given various weights depending on the scenario. Most of the other studies contain modifications to Proportional Fairness, using various weight functions to optimize fairness in the bandwidth allocation. Furthermore, [52] discussed the related works in this field and presented EXP and LOG rules as reliable approaches for delay-sensitive traffic in LTE downlink scheduling.

None of these approaches can guarantee the delay requirement of the flows, which is an important parameter to be considered in user satisfaction. In general, it is not desirable to satisfy the average delay requirement, but to enforce the upper bound delay tolerance guarantee. Furthermore, the aforementioned studies consider TTI by TTI scheduling and they do not present any plan to improve the system performance. Finally, they do not consider that some flows with data close to its deadlines may experience a poor channel quality, which would violate delay requirements. This problem is addressed in [53] by proposing two level scheduling algorithm. The upper level considers an approach based on discrete-time linear control theory and lower level utilize a proportional fairness scheduler. Also , some comprehensive surveys are provided in [54]. In [55], an opportunistic feedback scheme is proposed that combines techniques from queue/weight opportunistic scheduling to improve users' delay performance.

3.2 QoS-Aware MAC Scheduler and Offloading

The MAC scheduler design and algorithms used to allocate the radio resources has a critical impact on the eNodeB and overall LTE performance; therefore, the benefits of a QoS-aware MAC scheduler for the operators and vendors is obvious. In LTE, all services have delay requirements to be met. It is assumed that all the packets include time-stamps, so the scheduler monitors the service delays and channel conditions to allocate the resources.

Typically, the data rate that is available on the radio interface is smaller than the data rate available on the network interface. Thus, when the data rate of a given service is higher than the data rate provided by the LTE radio interface, this leads to buffering in the UE and in the eNodeB. This buffering allows some freedom to the scheduler in the MAC layer. In other words, it allows the scheduler to vary the instantaneous data rate at the physical layer, in order to adapt to the current radio channel conditions. However, when the data rate provided by the application exceeds the data rate provided by the radio interface for a long period, large amounts of buffered data can result in an excessive delay for delay sensitive applications.

Therefore, to avoid increasing the delay in the scheduler for the delay sensitive services, which results in packet loss and retransmission, a QoS-aware MAC scheduler algorithm to diverse a mix of traffic, including delay sensitive, is proposed. However,

as it is not feasible only to rely on the scheduler to provide the required QoS under heavy traffic loads, it is proposed to combine this scheduler with an intelligent offloading mechanism in an integrated LTE and Wi-Fi network to provide better performance and fairness through offloading as well as improved scheduling algorithm. The results demonstrate the effectiveness of the approach which makes real time decisions by considering a more consistent traffic prioritization when selecting and offloading the data traffic, based on the network conditions and QoS required by the user.

3.3 Problem Statement

The well-known proportional fairness (PF) scheduling algorithm has been extensively used in the literature; However, the fairness is only considered in terms of the required resources by the user and not the network resources' utilization. Therefore, the utility function of PF scheduler is modified to a function such that take into account both maximum achievable throughput associated with the network fairness and user required resources at the same time. The proposed novel scheduling mechanism targets real-time as well as non-real time traffic in the LTE downlink. The proposed algorithm shown in figure 3.4 try to assign radio resources to UE by taking into account the channel condition, the maximum tolerable delay and incoming data traffic rate.

Although the scheduler proposed here, is designed to serve both real-time and non-real time traffic, considering the delay weight will penalize the non-real time traffic or some packets which experience larger delay than their maximum delay tolerance, therefore the aim is to maximize the total end-to-end throughput of the system considering the flows QoS requirements. To achieve this, the QoS-aware scheduler is combined with smart offloading mechanism to offload such traffics on complimentary access networks like Wi-Fi. The two stages of this problem is detailed in this section.

3.3.1 QoS-Aware Flow Scheduling

Let us denote by U set of M active users, such that $U = \{u_1, u_2, ...u_m, ..., u_M\}$ carry a set of traffic flows belongs to set $F = \{f_1, f_2, ...f_i, ..., f_I\}$ with traffic type T belongs to $T = \{1, 2, 3, ..., J\}$. All traffic flows share the wireless channel and packets waiting for transmission are stored in a queue in the logical buffer associated to each traffic type for each user. The scheduler evaluates the transmission needs of all these queues at the beginning of each TTI. Furthermore, scheduler updates $d_{ij}^{curr}(n)$, the current achievable data rate for flow i with traffic type j at the n^{th} TTI, $d_{ij}^{ave}(n)$,the average data rate for

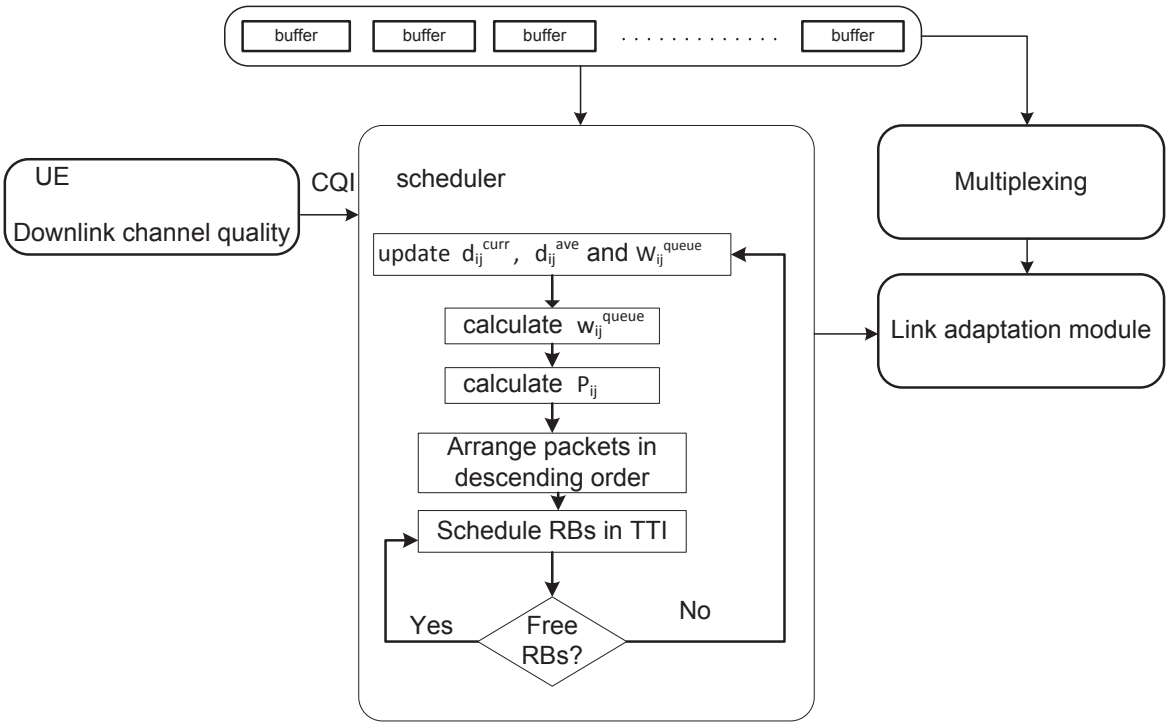

Figure 3.4: LTE Resource Block

the traffic type j of f_i up to the n^{th} TTI. The QoS parameter is considered in terms of weight, e.g. $w_{ij}^{d}(n)$ to denote the delay weight factor for flow i with traffic type j. The priority value for scheduling each user flows is calculated as follows:

$$P_{ij}(n) = \frac{d_{ij}^{curr}(n)}{d_{ij}^{ave}(n)} \times w_{ij}^{d}(n) \tag{3.1}$$

The average data rate is given by:

$$d_{ij}^{ave}(n+1) = \alpha \times d_{ij}^{ave}(n) + (1-\alpha) \times \beta_{ij} \times d_{ij}^{curr}(n) \tag{3.2}$$

α is the weight considered for data received in the previous TTIs and defines a trade-off between fairness and throughput. The higher the α is, the higher fairness is achieved among users.

$$\beta_{ij} = \begin{cases} 1 & \text{if packets of traffic type } j \text{ of user } i \text{ are selected for TTI n} \\ 0 & \text{Otherwise} \end{cases} \tag{3.3}$$

To support QoS, we consider a weight factor for delay which is defined as follows:

$$w_{ij}^{d}(n) = \exp(\frac{W_{ij}^{queue}(n)}{D_{j}^{max}} - 1) \tag{3.4}$$

$W_{ij}^{queue}(n)$ is the queue waiting time of HOL packet for flow i with traffic type j and D_{j}^{max} is the maximum delay budget of flow with the type of j. Upon arrival of the flow at the n^{th} TTI, $W_{ij}^{queue}(n)$ is estimated based on the flow size and data rate. This leads to find the w_{ij}^{d} and $P_{ij}(n)$ consequently. In each TTI, the scheduler arranges the users traffic based on the P_{ij} in descending order and allocates the RBs to the flows . Additionally, it will serve users with the highest achievable bit rate to maximize throughput until a difference in the amount of data has been served between users is obtained. Upon such a trigger, the users with significant lower received data take precedence over others.

The proposed scheduler targets mixed of real-time and non-real time traffic and considers three factors to calculate the priority values for flow i with traffic type j,(1) the current achievable data rate for each user is calculated for every TTI by link adaptation module in each sub channel using the channel quality information received from UE in the previous TTI. In LTE, the channel quality decides the modulation and coding scheme (MCS). The better channel quality, the higher MCS is used, therefore more symbols are carried by RB and higher data rate is achieved which also reduces the packet delay and improve the system performance. However, prioritize UEs with best channel condition is not fair for UEs under bad channel condition, (2) $d_{ij}^{ave}(n)$ introduces fairness among UE by taking into account the amount of data each UE received in the previous TTIs, (3)the delay weight factor incorporates the effect of upper bound delay that can be tolerated by each type of traffic. Also in this algorithm, if a user experience a bad channel condition and cannot be scheduled in a given TTI, this scheduler takes into account the pending data in the user buffer and schedule them in the next TTI.

3.3.2 Offloading Mechanism

The problem is defined such that the Wi-Fi access points try to accommodate best effort traffic and real-time traffic which will be affected by the proposed scheduling algorithm of the MAC scheduler. In the definition of this problem, it is assumed that a single flow can not be split between two different access points. In this model, Wi-Fis are considered as the knapsacks with defined capacity and the flows are the items. Consider that not all flows are in the range of all Wi-Fis, a set of flows per Wi-Fi is

defined, these flows are the items that are needed to fit in the knapsack.

We assume each traffic flow is characterized with the maximum delay budget, D_j^{max}, and minimum data rate required, d_j^{req}. Throughput at the access point c , Λ_c, can be defined as follows ,

$$\Lambda_c = \sum_{f_i \in p} Th_{f_i} \quad \forall p \in \{1, ..., P\} \tag{3.5}$$

We define the problem such that the objective is to maximize total end-to-end throughput [56].

On the other hand, offloading some of the flows to the Wi-Fi access point may be violating their QoS constraint, and thus they will remain in service of the LTE access point. The QoS parameter of flow i and the QoS boundary value are denoted by Q_i and Q_0. Although the same constraint is applied to all the flows in this formulation, it can easily be extended to different QoS constraint for each one of the flows. Also, assuming the total capacity of the access point c is C_c and the percentage of the total capacity usage which operator is interested to be occupied for each access network in the scenario is λ, there is an additional constraint on each of the access points's capacity. Therefore, the decision is constrained to,

$$Q_i \leq Q_0 \qquad \forall i \in \{1, ..., I\} \tag{3.6}$$

$$\Lambda_c \leq \lambda C_c \qquad \forall c \in \{1, ..., R\} \tag{3.7}$$

Thus the problem is formulated as follows:

$$\text{Maximize} \sum_{i=1}^{I} \sum_{c=1}^{C} x_{ic}\, Th_i \tag{3.8}$$

$$\text{Subject to} \sum_{i=1}^{I} \sum_{c=1}^{C} w_i\, x_{ic} \leq C_c \tag{3.9}$$

$$x_{ic} \in \{0, 1\}, \qquad \forall (i, c) \tag{3.10}$$

$$D_{ij} \leq D_j^{max} \tag{3.11}$$

Constraints (3.10),(3.11) states that a flow i can only be admitted if there is enough capacity in any of the access point and the flow is either admitted to Wi-Fi or LTE.

Constraints (3.11) considers data rate as a QoS performance metric and states that

the data rate requirements of the flow i with traffic type of j, D_{ij}, should be satisfied by the access network which is admitted to.

3.4 Performance Evaluation

To study the effectiveness of the proposed resource allocation in heterogenous network comprising of macro-cell and Wi-Fi access network, the impact on the Quality of Experience (QoE) perceived by end users for real-time flows have been analyzed in terms of meeting the delay requirement. Simulation results demonstrate that the proposed resource allocation scheme is able to respect the QoS constraints of real-time flows i.e. delay budget, guarantying the best QoE. Also, the proposed scheme is able to provide better performance for a mixed traffic.

3.4.1 Simulation Scenario

Downlink resource allocation in a multi-access network scenario consists of a macro-cell and Wi-Fi access is investigated here. A single eNodeB sitting at the center of each cell covers certain area of a single macro cell with the radius of 1 Km and performs the resource allocation and scheduling in every TTI. The cell also is covered by a number of randomly distributed Wi-Fi access networks, which their coverage area does not overlap each other, i.e. each mobile user can only get service via a single Wi-Fi access point at any location. One wireless access point with a power of 100 mW is deployed at the center of each hotspot. That is why co-channel interference between Wi-Fis is not included. Further it is assumed, there are enough Wi-Fi resources to maintain the current Wi-Fi data rate; hence data rate that Wi-Fi access point can offer is independent of its traffic load in the network. As Wi-Fi and LTE networks operate on different frequency bands, there is no resource partitioning or interference between these access networks. Also interference between adjacent Wi-Fi access points is not considered (relying on the Wi-Fi planning strategy), load balancing across them, and power control across overlapping access networks. Wi-Fi user is assigned a link rate, which yields the shortest expected time of transmission, assuming geometrical distribution for the number of trials. A link rate is also assigned for the acknowledgment packet. The Maximum possible link bitrate in 802.11n network on 20 MHz channel bandwidth and QPSK coding rate is 43 Mbps and it uses the 2.4 GHz frequency band. The difference with the LTE model is that active radio link bitrate of each user is scaled by a mapped value of the utilization factor of the access point that the user is

Table 3.1: Performance Requirement by Service Category

Service ID	Service Type	Rate (kbps)	Delay (ms)
1	VoIP	21-64	50-100
2	Video	500-700	100-200
3	Internet	50-600	300-600

connected to. The utilization factor of an access point is obtained by summing up the link utilization of the users connected to that access point. Furthermore it is assumed that the users are uniformly distributed in the coverage area of the macro cell. There are variable number of users exist in the coverage area of the macro-cell in each round of simulation and the mobility of each user traveling the cell is defined by the random way-point model [57]. Finally the best effort flows are considered as infinite buffer. Each simulation lasts 10 s and all simulation results are averaged over 5 simulations. A flow is defined as a bit stream generated by the application layer. Packets waiting for transmission are stored in a queue associated to the user buffer for each type of traffic. Queues are assumed as infinite buffer size to ignore the buffer overflow effects. The services and their requirements are summarized in Table 5.1.

3.4.2 System Model for LTE

The main simulation parameters are detailed in Table 3.2. Packet scheduling is considered based on the parameters for the downlink of the LTE-E-UTRAN suggested by 3GPP specification [58]. In the LTE MAC scheduler, the objective is to maximize the throughput by increasing the instantaneous achievable data rate while maintaining fairness in resource allocation. Available 3 MHz bandwidth has been chosen for the system which leads to a total of 15 pairs of contiguous available RBs in every TTI. For the channel model, a propagation loss model has been considered [58]. The contributing factors affecting the received power at the UEs in such a model are free space path loss, shadowing and fast fading. The free path loss between two nodes is determined using standard radio propagation models which considers the loss, L, as a function of the distance between eNodeB and the user in Km, d, on the form as defined by the equation given below.

$$P_L = 128.1 + 37.6 * \log d \tag{3.12}$$

Shadowing is modeled as a log-normal random variable with zero mean and standard

deviation of 8dB. Therefore the SINR at the UEs is calculated as follows:

$$SINR_{ij} = \frac{P_{rij}}{FN_0B} \tag{3.13}$$

Where P_{rij} is the received power measured at UE i through jth subchannel. F, N_0, B are noise figure, the noise spectral efficiency and the bandwidth of the RB respectively. UE estimates the SINR of the received signal for all available downlink sub-channels every TTI and then cross checked with a quantized SINR level to map onto the a 15-element in the CQI array and then transmitted to the serving eNodeB using uplink channel.The CQI values are obtained as a quantized value of the estimated SINR to guarantee a Block Error Rate (BLER) equal to 10%. The mapping between SINR and CQI is obtained from BLER-SINR curves similar to [59]. The UEs will determine whether or not a packet is received in error based on the SINR being above the threshold [60]. The required number of RBs for each transmission is calculated based on the chosen type of service which will be calculated based on the data rate. In the scheduler, firstly, a list of available RBs is created and updated each TTI and then based on the full-band periodic CQI reports received in the cell, eNodeB map each CQI to a MCs using the Exponential Effective SINR Mapping method [21] which will also have a corresponding spectral efficiency associated with it. The spectral efficiency is defined to calculate the instantaneous achievable data rate using a known number of contiguous RBs and the number of symbols used. Therefore, the utility function is computed for all UEs over all sub-channels and number of RBs from the previously updated list of available RBs are optimally allocated to each UEs. After the scheduling process, an average SINR over the allocated RBS is estimated by eNodeB from the latest received sub-channel CQIs from UEs and then mapped to the best MCS for each transmission [34].

3.4.3 Simulation Methodology

The important parameter considered here for scheduling is the target delay for real-time flows. In real-time services such as VoIP and Video, the maximum allowed end-to-end delay is considered in the range of 100-200 ms [61]. Therefore, the target delay for last hop which is in this scenario is between UEs and eNodeB should be considered less than these values. Furthermore, the delay target is set to 50-150 ms in the simulation which is less than what is suggested by 3GPP specifications [62]. In each TTI, the transmission time required for each flow is calculated based on its data rate. Also, based on the bits to transmit for this flow in the buffer, the expected scheduler transmission

Table 3.2: Simulation Parameters

Network	Parameter	Value)
Cellular Network	Downlink Transmission bandwidth	3 MHz
	Carrier Frequency	2 GHz
	Number of available PRB	15
	Number of available subcarrier per PRB	12
	Subcarrier spacing	15 KHZ
	Number of symbols per PRB	7 (normal cyclic prefix)
	PRB bandwidth	180 KHZ
	TTI duration	1 ms
	eNodeB transmission power	3 mW
	Modulation schemes	QPSK,16QAM and 64QAM with all avaiable coding rate
	Cell Radius	1 Km
	CQI	Periodic and full bandwidth
	Chanel State Information (CSI)	Known and non-varying
Wi-Fi Network (802.11g)	Range	50 m
	Number of users attached	40
	Transmission power	100 mW
	Simulation duration	1 s

time for this flow is computed. Hence, average waiting time in the eNodeB buffer is preemptively computed for each type of traffic. The delay weight factor is computed and feeded to the utility function. The MAC scheduler performs resource allocation and allocates RBs optimally. This mechanism ensures that real-time services such as VoIP and Video receiving their QoS requirement. However, this mechanism is penalizing the best effort traffic as they do not have any QoS requirement. This issue is addressed through offloading the best effort traffic to Wi-Fi.

3.4.4 Numerical results

The performance of proposed resource allocation in multi-access network scenario is examined by varying number of users, their distance and the target delays for real-time flows. The examined metrics here include **Aggregate Throughput** and **Number of Completed** with respect to their delay requirement for each service. Since the best effort flows do not have any strict QoS constraints, the aggregate throughput is considered to compare resource allocation strategy.

The throughput increases with the number of UEs, due to the higher load on the network. However, UES mobility pattern can result in sudden changes of the channel quality in two consecutive TTI and degregade the throughput. Figures 3.5, 3.6, 3.7 show the aggregate throughput for all users' voice, video, and data services. Offloading

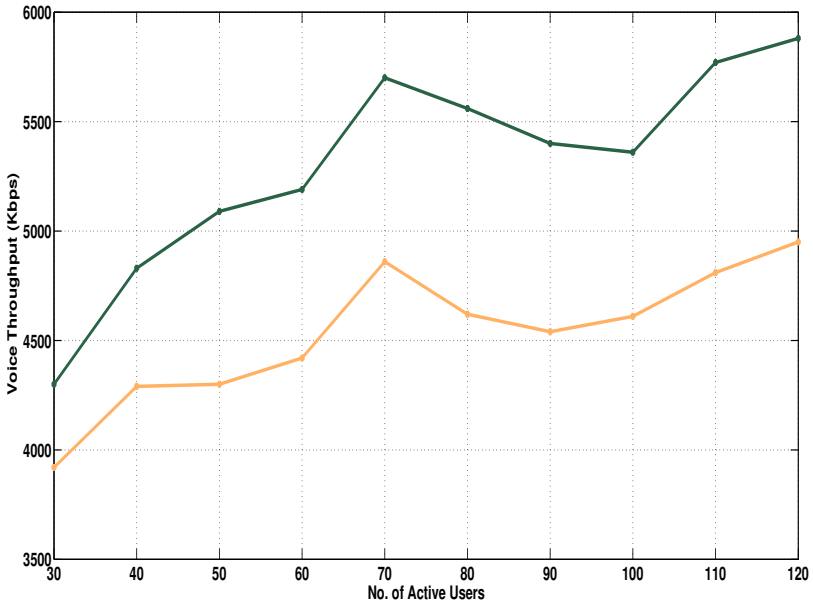

Figure 3.5: Aggregate Voice Throughput

best effort data on Wi-Fi network result in more available resources for voice and video. As it can be seen, the Aggregate throughput of voice and video increased via QoS-aware scheduling while the best effort achieve maximum throughput through offloading.

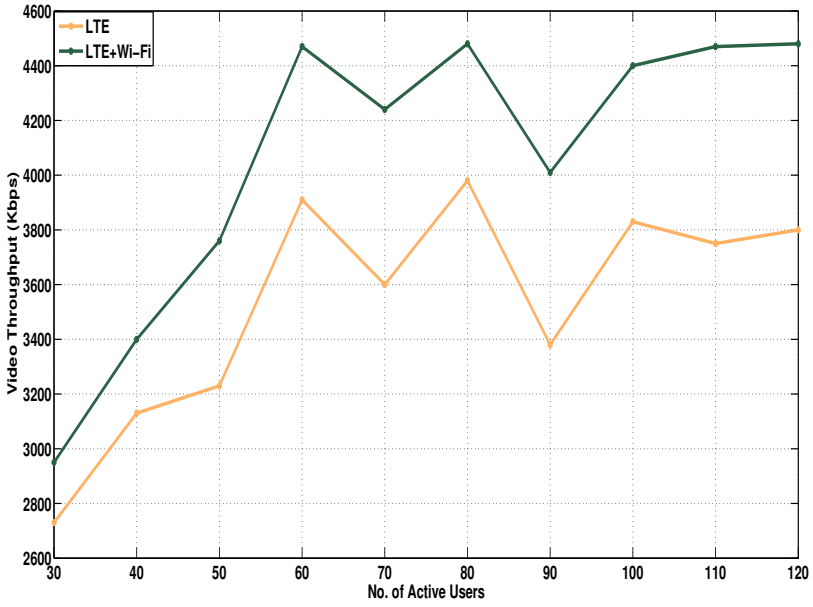

Figure 3.6: Aggregate Video Throughput

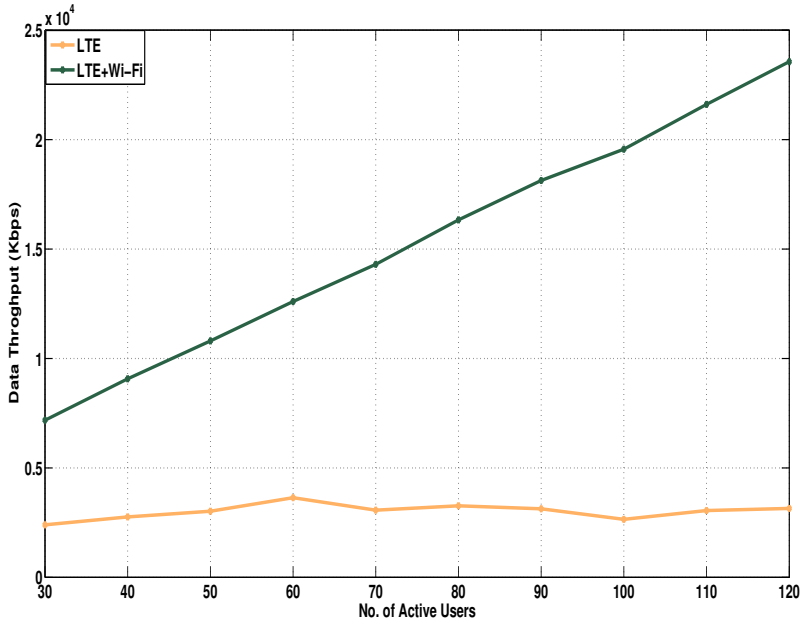

Figure 3.7: Aggregate Data Throughput

Figures 3.8, 3.9, 3.10 depict number of flows have been served. For voice and video, these figures quantify number of flows completed within their delay requirement. Data does not have any delay requirement and can be served by both network. The significant increase in the number of data flows completed proves that offloading compensate the LTE scheduler mechanism to prioritize real-time traffic.

3.5 Conclusion

QoS-aware scheduling algorithm combined with offloading mechanism is designed with the mixture of optimization objectives for different traffic. This study supports a mixture of best effort and real-time traffic in realistic wireless scenarios, and ensures that real-time traffic receive its desired QoS while the best effort traffic which doesn't have any QoS requirement as such, achieves maximum possible throughput. To design such a scheduler, a utility function innovated to take into account the delay requirement of the services, fairness and achieving maximum throughput. The channel variation effect is exploited as an opportunity for the scheduler to tend to choose users with the best channel condition for transmission in every TTI. However, it should provide fairness to other users. Here fairness is addressed by taking in to account the aggregate average data received by the user. A weight factor in the utility function is working

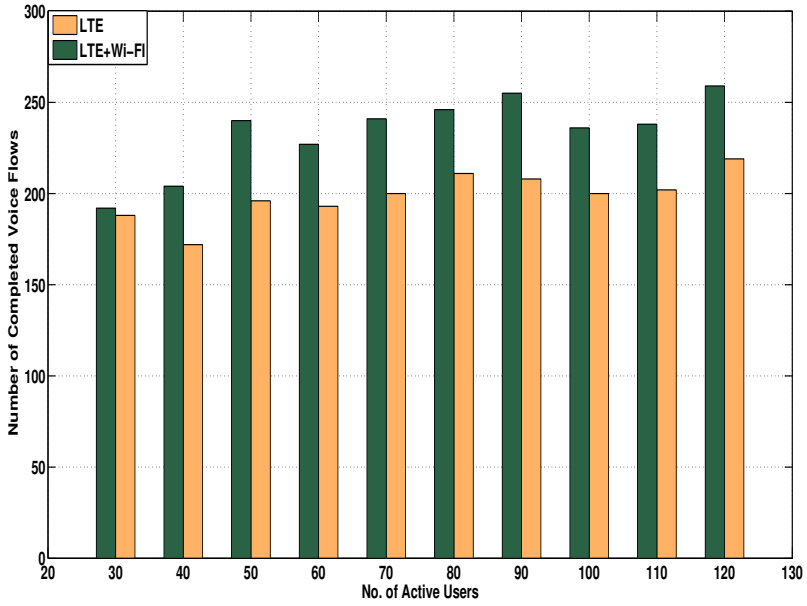

Figure 3.8: Number of Completed Voice Flows within Delay Requirement

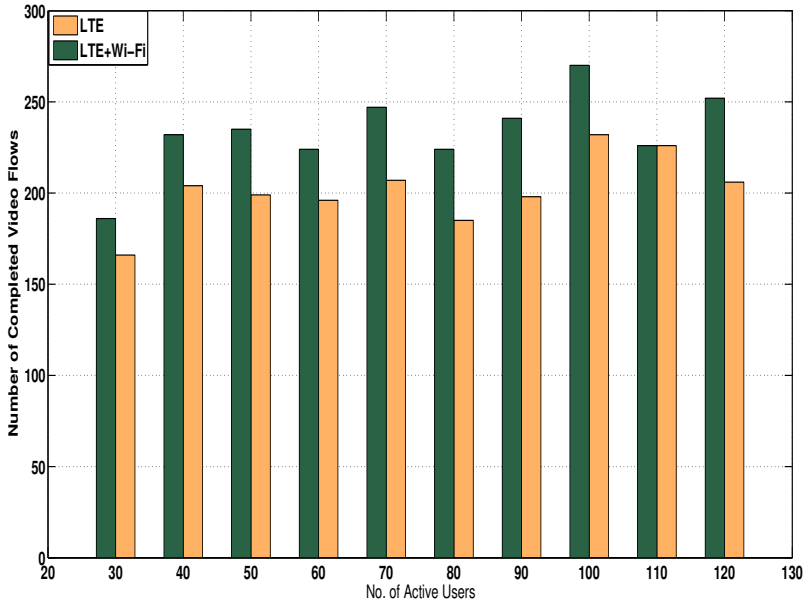

Figure 3.9: Number of Completed Video Flows within Delay Requirement

in the favor of real-time traffic and against the best effort services. To address the consequence of this effect, the proposed mechanism is combined with offloading, to provide service to best effort traffic. The proposed scheme is simulated and numerical

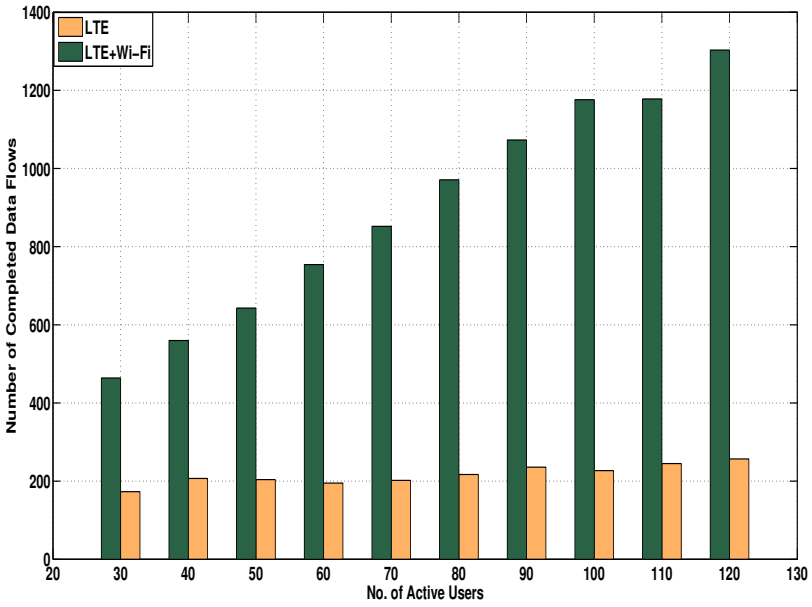

Figure 3.10: Number of Completed Data Flows

results prove the effectiveness of such an algorithm.

Chapter 4

Policy-Based Mobile Data Offloading

4.1 Introduction

Beside the evolution of beyond 3G systems of heterogenous networks including many radio access technologies such as 2G/3G/EDGE, UMTS/HSDPA, LTE, 802.11 and 802.16 which are reviewed in Chapter 2, user equipments (UEs) also have multi-mode capabilities and can support multi-homing. Additionally, mobile terminals (MTs) and base station can both support dynamic spectrum allocation [63]. This radio context prepared the offloading platform, but still there is a need to define standards to deal with advanced radio resource management protocols which can be adapted to the present and future heterogenous networks. On this note, in this chapter policy based offloading framework is proposed where offloading decisions are based on a set of rules that consider both network conditions and operator objectives/strategies. The network centric approach considers network and environment condition in the decision making, therefore guarantees a better stability and robustness of the system but introduces higher signalling load. Policy based network management schemes in the literature, are based on the network centric approach, where all the decisions are made by the network. From the other side, user centric approach alleviates the signalling overload and introduce better adaptation to the user profile, therefore enhances the QoS; however it violates the overall system stability. A better approach is hybrid decision making which is a trade-off between the system stability and QoS enhancement for the user. Therefore, a hybrid policy where decision making is shared between both the user and the network is introduced. The network instructs the user about the available options and orients the user's decision by sending rules and constraints. User chooses

the most appropriate option to their QoS objectives within the set of rules allowed by the network. The main contributions of this chapter are summarized as follows:

1. The state of the art policy-based decision making is presented.
2. A novel mechanism based on principles of autonomic networking is proposed to incorporate the effect of real-time network conditions into offloading policies.
3. To evaluate the performance of the proposed framework, cost-function approach is taken and network centric, user centric and hybrid policies are discussed.
4. Detailed simulations are performed to validate the effectiveness of these policies in real networks. The offloading efficiency (OE) and blocking ratio (BR) are compared for the mentioned framework.

The rest of the chapter is organized as follows. Section 4.2 explains the proposed policy based offloading framework along with the proposed architecture for shared decision making between the network and the user. In Section 4.3, performance evaluation of these policies is carried out. Finally the conclusions are drawn in Section 4.4.

4.2 Policy Based Offloading Framework

In this section, the proposed policy based offloading framework is described. The policies are based on a cost function which derives the decision in case of user centric, network centric and hybrid approaches.

4.2.1 User Centric Offloading Policies

In the user centric approach, the user performs different inter-system measurements and makes a network selection decision considering the terminal capabilities and service requirements. The users equipments are considered to be multimode. This approach is used in cognitive radio based system where the UEs are intelligent and can learn from the previous experiences [64]. The mobile terminals takes decision only based on its own measurement and QoS requirement without any precise information about the network state. The cost function in this case can be defined as follows.

$$C_n = w_s S_n + w_\varphi \varphi_n, \qquad for \;\; n = 1, 2, ...M \tag{4.1}$$

where C_n is the cost to use network n, S_n is the normalized relative received signal strength from network n, φ_n is access fee to use network n, w_s $(0 < w_s < 1)$ and

w_φ $(0 < w_\varphi < 1)$ are the weights that indicate the impact of S_n and φ_n on the user selection respectively. The access fee, φ_n

$$\varphi_n = \frac{\rho_{max} - \rho_i}{\rho_{max}} \tag{4.2}$$

where ρ_{max} is the maximum access fee that the user is willing to pay and $\rho_i$is the access fee to use network n. Therefore, the network with the cheaper access fee has a larger cost. The constraint between w_s and w_φ is given by

$$w_s + w_\varphi = 1 \tag{4.3}$$

stronger received signal indicates better signal quality, and a user prefers to select a network that provide higher received signal strength, it is not easy to compare different types of access networks in terms of received signal strength due to different maximum transmission powers and receiver thresholds. Therefore, the use of relative received signal strength to compare different types of networks in terms of received signal strength is proposed in [65]. When user has more than one available network to select, the costs of all available networks is calculated by 4.1. Then the user selects the network with the largest cost function, which indicates a network with cheapest access fee and highest relative received signal strength. A special case of this scenario is when $w_\varphi = 0$, i.e. the received signal strength quality is the only factor to select the network. It should be noted, although the cost function comprised of just two components, it is easy to extend to involve more factors such as the battery consumption. However, the additional factors should be added after proper normalization. It should be noted that the user centric network selection is a type of unmanaged offloading (and hence the name user centric offloading) in which the user data is transparently moved to another network, whenever that network is detected. The operator can deploy such an offloading solution by simply placing an application in the terminals e.g., in case of Wi-Fi, the application will turn on the Wi-Fi interface whenever a Wi-Fi signal is detected. The user data in this case would be transparently moved on to the Wi-Fi network. The user centric offloading has certain weaknesses. As the decision is made only by the user based on its own measurements and capabilities, therefore the user does not have a precise idea of overall network conditions. For example, the user can select the Wi-Fi network based on strong signal strength. However he may experience a poor service if the Wi-Fi network is congested in the backhaul. From operator's perspective, the user centric offloading leads to losing the visibility and hence control of the users. Moreover, the operator cannot deliver any subscribed content leading to

a potential loss of revenue.

4.2.2 Network Centric Offloading Policies

The network centric offloading approach, a central entity controls all the decisions between heterogenous networks after performing different inter system measurements and collecting information from different access networks. In this approach, a central entity in the core network controls all the decisions between heterogenous networks. This central entity is advised all the entities accordingly and these entities are bound to execute the decisions. There is a tight coupling between central entity and the radio resource management entities. From network perspective, parameters such as bandwidth utilization, traffic load, and congestion are important for making an offloading decision. A simple realization of network centric offloading policy is a load balancing scenario among multiple access networks. The respective cost function in this case is given as follows.

$$C_n = w_g G_n \quad for \;\; n = 1, 2, ...M \tag{4.4}$$

where G_n is the normalized utilization of network n, and w_g $(0 < w_g < 1)$ is its associated weighting factor. By defining different load classes such as low, medium and high, a load balancing policy will be triggered whenever one of the access networks is overloaded which means operating under high traffic load, and hence has a higher cost value. This policy can be sent to terminals requiring a mandatory action i.e., moving from the higher utilized network. The terminals on reception of the policy will take the mandatory action and connect to other, less utilized networks. Moreover, the policy can also be sent to a selective group of users, e.g., the network would like to retain the higher priority users and thus only lower priority users will affected. Compared to user centric offloading, the network centric offloading has overall information about the system state, therefore the overall decision is more robust. However it has an associated penalty in terms of extra signaling load that arises due to exchange of various inter-system measurements.

4.2.3 Hybrid Offloading Polices

In the hybrid offloading policies, the offloading decision is shared between the user and the network. This provides a tradeoff in terms of decision complexity and robustness between a decision totally controlled by the network and user respectively. The network manages the overall policy framework by deriving the policies considering the operator

strategies and network conditions. These policies are periodically sent to users which contribute to achieving the network objectives by making an optimum offloading decision and maximizing their own performance by combining network information with their own measurements and service requirements. This hybrid approach results in enhancing the overall system performance through optimized offloading decision while at the same time the user experience is improved as well. In order to realize this hybrid approach, an architecture and mechanism which is based on principles of self-management and autonomic networking is proposed which is firstly discussed before carrying on to use the cost function to model this problem.

Hybrid Offloading Architecture

The concept of self-management is derived from IBM's vision of autonomic computing which is based on four aspects; self-configuration, self-optimization, self-healing and self-protection [66]. This concept of autonomic computing has been further extended to autonomic networking. In order to create self-managed systems, the aspects of self-organization and self-awareness are also necessary. Autonomic operation of a system (or network) is achieved using feedback or control cycles. In a generalized form this cycle, also known as MDE cycle, involves interactive feedback steps for collecting inputs from environment and involved elements (Monitoring), reasoning and learning based on certain algorithms and available knowledge (Decision Making) and invoking actions for achieving desired goals in the system (Execution) [67, 68]. The underlying architecture is illustrated by considering a tightly coupled cellular/Wi-Fi system where both share the same core network as shown in Figure 4.1. The CNME (Centralized Network Management Entity) is an autonomic entity, located in the core network, having an end-to-end visibility of the entire network. Similarly LNMEs (Local Network Management Entities) are autonomic entities (or agents [66]) have local visibility of elements such as Wi-Fi access points and cellular base stations. The mechanism is described as follows. During the Monitoring phase, different QoS parameters are monitored. The LNMEs for cellular and Wi-Fi networks gather these necessary parameters (as shown in Figure 4.1) and periodically report to the CNME. For the Wi-Fi network, important parameters are the number of users attached, bandwidth utilization etc. Similarly for the cellular network, the availability and load information of base stations are important. During the Decision Making phase, the CNME interprets this information in order to have a real time visibility of network conditions. It then derives the most appropriate set of policies subject to operator strategies and network conditions. In the Execution phase, these policies are sent to the user via the cellular network (using different physical

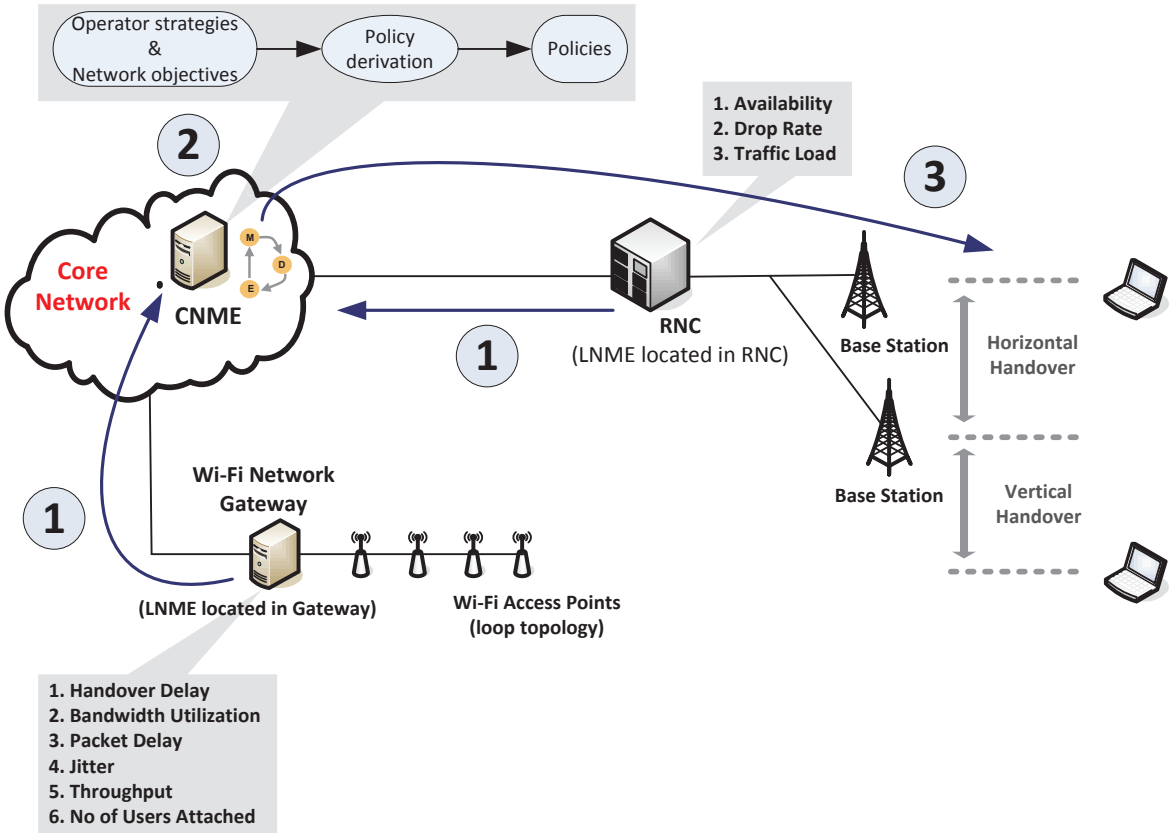

Figure 4.1: Policy Based Offloading and Mechanism(step1,2,3 correspond to Monitoring, Decision Making and Execution respectively

channels). In order to keep signaling load to a minimum, the frequency of sending these policies can be adjusted adaptively depending on significant changes in the network conditions or states. The user will run a decision algorithm on the device which uses the information sent by the network and makes the final decision based on its own profile such as the requested service, location, channel condition and previous experiences [69]. Alternatively, the network can take more proactive role and enforce the user to choose specific access network. In this scheme, the network only provide a list of operations to each user but does not decide which operation should be performed which reduces downlink signalling. The uplink signalling is also reduced as the user does not need to send their information to the network.

Hybrid Offloading Gains

The shared decision between user and network aims to enhance the overall network performance. The main idea is allowing the user to make decisions autonomously con-

sidering the constraints and policy defined by the network. The network manages the overall system through policies, but is not in charge of full decision making procedure.

- Gain for network : The dynamic policy derivation intends to optimize the system performance by taking into account the operator strategies and real-time network conditions. On the other hand, user will benefit from flexibility in its connection within the framework of the policies instructed by the network. However given a policy, a user can choose any of the access networks with a certain probability which introduces a level of uncertainty on the overall behavior of the users. Therefore, the network should predict the reaction of the system to its policies, e.g. a prior knowledge of the user reaction to a given policy affects the system behavior prediction and policy derivation. For example a simple policy in the network compromising of cellular and Wi-Fi access networks is to connect to cellular or Wi-Fi but simultaneous connection is not allowed. After having activated this policy several times, the network observes that in average user is connected to the Wi-Fi with the probability of 60%. Then network can refine its policy if it wants to change this rate e.g. by adding extra constraint such as just high data rate services can connect to Wi-Fi.

- Gain for the user: The user benefits from the flexibility offered in the hybrid mechanism. Although some policies might be very restrictive but also there is a margin for user to setup its preferred connection. Through adopting the cognitive mechanism from [64], learning the user habit dynamically from previous experiences has added value, due to the fact that the user decision can be adopted by its requirement.

Hybrid Offloading Model

The user decision will be based on a cost function which is defined here. As the cost function in this hybrid offloading should reflect the incorporation of network conditions in the user's decision making process, therefore a network elimination factor given by E_s^n is introduced for a network n, which expresses the inability of particular network to provide a specific service s at a certain time. Thus, E_s^n can have a value of 0 or 1. The QoS factor Q_s^n depicts the user preferences and requirement for a specific service. By defining thresholds of different QoS indicators for commonly used services, such as voice, web browsing, video streaming etc., the network can send recommendation for different networks whether they can support a certain service with acceptable QoS. For example, if the Wi-Fi network has significant packet delay due to congestion in the

backhaul, and thus cannot support a Voice over IP (VoIP) call, the network can send the information about packet delay thresholds to the user along with the recommendation that VoIP cannot be supported at this time. The potential terminal upon receiving this information compares the packet delay with the minimum required packet delay for VoIP and does not select Wi-Fi by setting E_s^n to 0.

$$C_n = \sum_s E_s^n Q_s^n \tag{4.5}$$

4.3 Performance Evaluation

4.3.1 System Model

A multi-access network scenario is considered whereby a single macro 3G base station provides coverage in a certain area. It is assumed that a number of Wi-Fi access points are randomly distributed in the entire coverage area of the macro 3G base station. Furthermore, it is assumed that the users are uniformly distributed in the coverage area of the macro 3G base station.

4.3.2 Traffic Models

For the cellular network, the traffic load (in terms of no. of users), *L(t)*varies as a percentage of busy hour load (BusyLoad) over a 24 hour period as shown in Figure 4.2, pertaining to traffic in a 3G network in London, UK, obtained via private interaction with Vodafone representatives in a research project.

For the Wi-Fi network, t$\lambda(t)$ represents the user association rate over a 24 hour period (expressed in seconds) and varies approximately according to a sinusoidal curve [70] as shown in Figure 4.3, which resembles closely the observations made in [71]. In addition to this, the active users on both cellular and Wi-Fi networks have poisson distribution with the mean taken from the traffic load at that time of the day i.e., *L(t)* for the cellular network and $\lambda(t)$ for the Wi-Fi network.

For example, for the cellular network, the probability that k number of active users is present at any time of the day t is given by 4.6 and the number of active users using a Wi-Fi access point is calculated by 4.7.

$$P(k,t)_{cellular} = \frac{L(t)^k e^{-L(t)}}{k!} \tag{4.6}$$

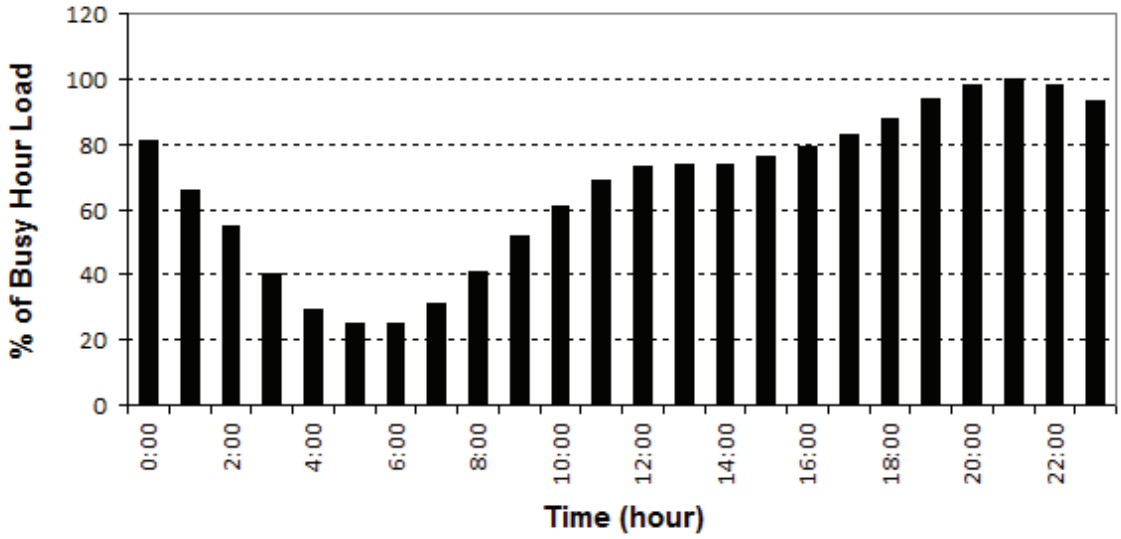

Figure 4.2: Cellular Busy Hour Load

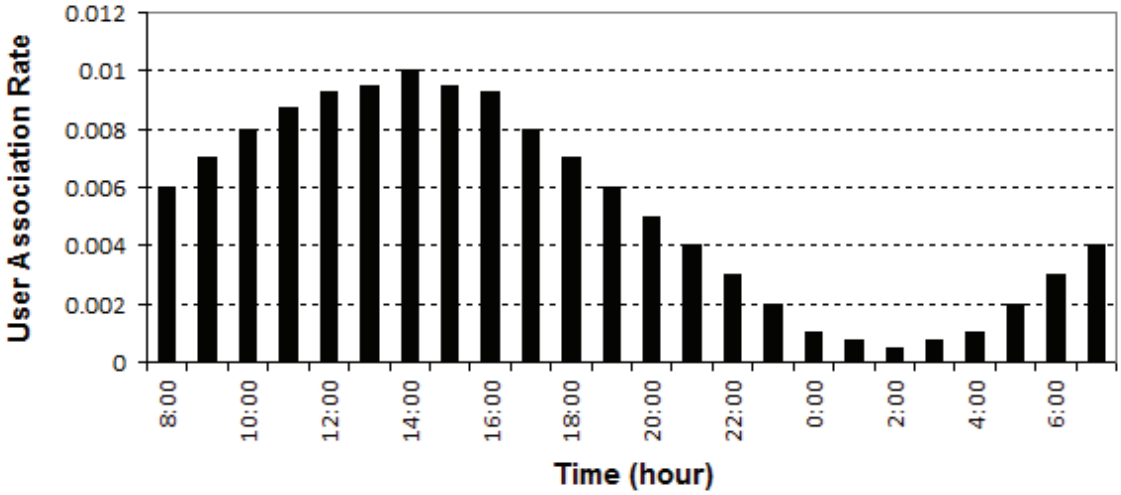

Figure 4.3: Wi-Fi User Association Rate

$$P(k,t)_{WiFi} = \frac{\lambda(t)^k e^{-\lambda(t)}}{k!} \tag{4.7}$$

For the cellular network, an important parameter is *Toffload*, which is the threshold (in terms of number of users in the cell) at which the offloading from cellular network

Table 4.1: Traffic Model Parameters

Traffic	Parameter	Value
HTTP	Reading Time (OFF duration)	Exponential Distribution Mean 30s
	Reading Time (OFF duration)	Exponential Distribution Mean 30s
	Parsing Time (OFF duration)	Exponential Distribution Mean 0.13s
	Main Object Size (ON duration)	Truncated Lognormal ($\sigma = 1.37, \mu = 8.35$) Mean 100B ,max=2MB
	Embedded Object Size (ON duration)	Truncated Lognormal ($\sigma = 2.36, \mu = 6.17$) Mean=50B ,max=2MB
	Number of Embedded Objects (ON duration)	Truncated Pareto ($\alpha = 1.1, k = 2, \max = 55$)
Video Streaming	Inter Arrival Time between Frames	Deterministic(10 fps) 100 ms
	Number of Packets/Slices in a Frame	Deteministic, 8
	Packet/Slice Size	Truncated Pareto (ON Duration)$\alpha = 1.2$, k=20B, Mean=50B, Max=125MB
	Inter Arrival Time between Packets/Slices (OFF duration)	Truncated Pareto $\alpha = 1.2$, k=2.5ms, Max=12.5ms, Min=6ms

begins. This threshold is carefully selected keeping in view the maximum capacity of the cellular network and acceptable utilization levels for optimum performance of all services. For the Wi-Fi network, it is important to take into account the probability of spatial and temporal coverage. The probability of spatial coverage, P_s, is defined as a fraction of macro 3G service area covered by Wi-Fi access points. It can be simply calculated by considering the number of Wi-Fi access points and their associated radii. The probability of temporal coverage, P_t, is defined as average time duration that a user stays the Wi-Fi coverage. An interesting study regarding the temporal Wi-Fi coverage can be found in [27].

The ON/OFF traffic model is adopted for the cellular network. Voice, web browsing and video streaming type services are assumed per user. The ON/OFF traffic model parameters are defined as given in [72], [73] and summarized in the Table 4.1. Similarly for the Wi-Fi network, the poisson model is used.

4.3.3 Simulation Methodology

Simulations have been performed for a UMTS/HSDPA type system where the available bandwidth per carrier is 5 MHz and a total of 2 carriers are available in the macro base station. For the Wi-Fi network, a total of 4 access points with a maximum of 40 attached users per access point is assumed. The maximum bandwidth available per access point is 54 Mbps. The numerical assessment cycles in outer loops through a 24 hour period in steps of t of one hour and uses the value of $L(t)$ or $\lambda(t)$ to parameterize 4.6 and 4.7. The inner loops it cycles through each possible value of number of active users, k, for cellular and Wi-Fi networks applying the user centric, network centric, and hybrid policies as discussed in Section 4.2. The results are averaged over all the simulations performed at each hour n the 24 hour period. The performance of different policies is compared in terms of Offloading Efficiency (OE) and Blocking Ratio (BR)

[27]. The OE is defined as the ratio of offloaded users,N_{off}, to the difference of total active users (on the cellular network), N_{total} and *Toffload* as given in 4.8. Similarly, BR is defined as the ratio of blocked users, $N_{blocked}$ (which could not be successfully offloaded) to the difference of total active users (on the cellular network) and *Toffload*.

$$OE = \frac{N_{off}}{N_{total} - Toffload} \tag{4.8}$$

$$BR = \frac{N_{blocked}}{N_{total} - Toffload} \tag{4.9}$$

Percentages for voice, web browsing and video streaming traffics are respectively selected as 25%, 35%, and 40%. The offloading threshold, *Toffload* is assumed to be 75% of the maximum capacity of cellular network (which for the above mentioned configuration comes out to be 72 users). The value for probability of spatial Wi-Fi coverage, Ps, is assumed to be 0.75 whereas the probability of temporal Wi-Fi coverage, Pt, is chosen as 0.65 for the day time (09:00 - 18:00) and 0.90 for the night time as most of the users have Wi-Fi coverage in homes [27].

4.3.4 Numerical Results

The results in Figures 4.4 and 4.5 respectively give the average OE(%) and BR(%) over a 24 hour period for different offloading policies. As shown by the results, the hybrid policy outperforms both network and user centric policies, by achieving an OE of up to 40% while keeping the BR to 10% under high load conditions. The OE results for user centric approach are not incorporated as it is unmanaged and not dependent on *Toffload* as in case of network and hybrid approaches. It should be noted that the OE under low load conditions is not significant primarily because the BusyLoad is below *Toffload*. However some figures can be seen as the active number of users is assumed to be Poisson distributed around the mean which is given by BusyLoad. The network centric approach achieves appreciable OE, however results in higher BR as the users are simply offloaded to Wi-Fi without considering the Wi-Fi environment. For simplicity the congestion on Wi-Fi is assumed in terms of minimum bandwidth requirement of 1 Mbps per user which should be satisfied. The user centric approach results in intermediate performance in terms of BR. This is because the user in the Wi-Fi coverage is assumed to make a network selection decision with a certain probability, unlike the network centric approach where it is bound to have a value of unity.

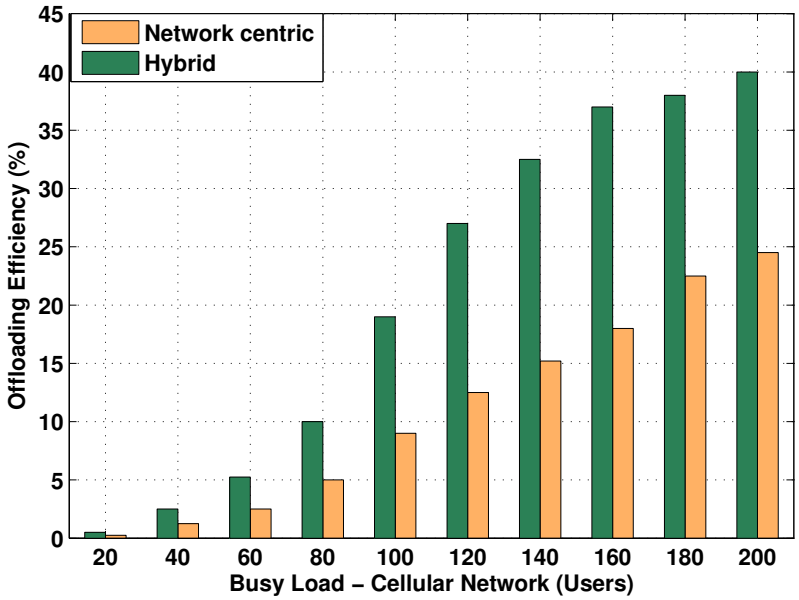

Figure 4.4: Offloading Efficiency over a 24 hour period against the BusyLoad.

4.4 Conclusion

Mobile data offloading is becoming a key industry segment due to the unprecedented pace at which data traffic is rising over the mobile networks. This chapter introduced a policy based offloading framework for efficiently offloading data traffic. With respect to offloading decision, a key aspect is decision sharing between the network and user which results not only optimum performance from network perspective but also improves the user experience. A novel mechanism for this hybrid decision making is presented. Simulation results point to a viable solution as significant OE of up to 40% can be achieved through such hybrid policies, under realistic traffic loads. Moreover, it also results in lesser BR, compared to user and network centric policies.

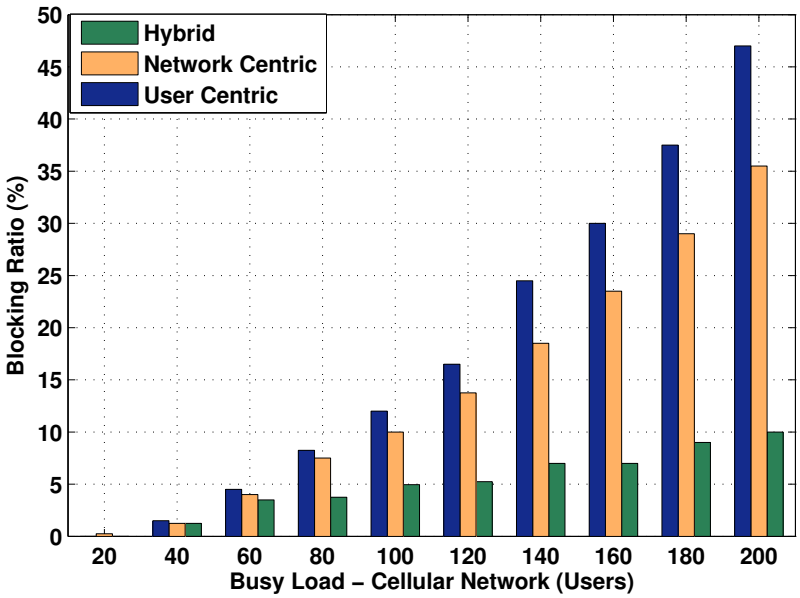

Figure 4.5: Blocking ratio over a 24 hour period against the BusyLoad.

Chapter 5

Programmable Policies for Mobile Data Offloading

Mobile phones will overtake PCs as the most common Internet access device worldwide in the near future. While mobile network operators will carry the bulk of Internet traffic, they face significant challenges in addressing the needs of increased traffic demand. As discussed in the previous chapters, scarcity of spectrum forces mobile operators to deploy smaller cells and utilizing of unlicensed spectrum such as Wi-Fi. Availability of built-in Wi-Fi on smartphones along with characteristic such as unlicensed spectrum makes Wi-Fi a natural solution for accommodating the increased traffic and maintaining the quality of users' connections. Furthermore, mobile networks must support various applications such as voice, streaming, as well as best-effort services on a single IP-based infrastructure. Each of these converged services has Quality of Service(QoS) requirements such as latency, packet loss and data rates, which must be obtained through efficient allocation of the wireless network resources and cannot be addressed only by provisioning the network. One of the main challenges in data offloading solution is making real-time decisions for offloading different users' flows and services/applications while taking condition of the available networks as well as the QoS needs of flows into account. In the previous chapter, this issue has been addressed through policy-based offloading framework, however applying these policies dynamically and scalability of such a framework is another issue to be covered in this chapter.

The mobile networks architecture as proposed in the 4G/LTE provides an easier management comparing to earlier technologies, by separating the management plane functionalities such as mobility management, policies and charging. Despite this, today's 4G/LTE architecture is not yet as flexible or programmable as can be. For

example, the 4G/LTE network is capable of offering QoS guarantees and differentiated services to the user, using Policy Charging and Rules Function (PCRF) nodes, while the Policy and Charging Control (PCC) ensures users' QoS for a particular subscription and service type. Introducing more new features, however, become increasingly complicated because the PCRF can not be dynamically programmed. Hence, this chapter intends to argue that the mobile communication architecture can be improved further by applying the principles of Software-Defined Networking (SDN) with providing logically centralized control of the overall infrastructure, and enabling programmability. The main contribution of this chapter is as follows:

1. Applying the abstraction of Software-defined Networking (SDN) in the mobile backhaul, couple the Network Resource Management (NRM) and Radio Resource Management (RRM) together to allow programmable and dynamic policy functions.

2. Moving the policy enforcement point to the place where congestion occurs, so that there is no need to transfer this dynamic information anywhere.

3. A programmable interface similar to what SDN offers is proposed to facilitates offloading mechanism by providing an end-to-end communication between network elements and pushing corresponding forwarding rules to the local elements (i.e., eNodeB, and P-GW).

4. Two methods of offloading policy functions are proposed where the rules and functions are derived by combining information from the RRM and PCRF applications.

5. Through investigations on the performance, significant increase in the number of offloaded flows and decrease in the dropped flows are achieved which points to network performance improvement.

The rest of chapter is organized as follows. Section 5.1 describes the LTE/Wi-Fi interworking architecture, and how programmable policy functions can modify this interworking and enable better QoS provisioning is elaborated in Section 5.2.2. Details of the policy functions and parameters that potentially affect offloading decisions are explained in Section 5.3. After describing the simulation scenarios in which the proposed offloading mechanism are examined, results are presented in Section 5.4. A summary of the proposed mechanism, as well as its overall results are presented in Section 5.5.

5.1 LTE and Wi-Fi Interworking Architecture

The goal in mobile data offloading is to redirect the selected traffic towards the lower cost radio access network dynamically, therefore 3GPP has been actively developed a new standard and architectures to support simultaneous use of different cellular access networks such as LTE, femtocells and non-3GPP access networks such as Wi-Fi [74, 75]. 3GPP standard differentiates two types of Wi-Fi access (also referred to as non-3GPP IP access):

- Untrusted: This type of access is introduced in the early stages of Wi-Fi specification in 3GPP release and refers to any type of Wi-Fi access that either does not support security, authentication and encryption or is not controlled by any operator such as public hotspot, home Wi-Fi.
- Trusted: Trusted non-3GPP IP access was introduced only for LTE in the later releases and refers to any operator's Wi-Fi with secure authentication method and over the air encryption. Although most of the offloading designs and scenarios are considering this type of access by default, it is only integrated into LTE's evolved packet core (EPC) and 3GPP does not support any integration to 2G or 3G core.

3GPP describes native integration of trusted and untrusted non-3GPP IP access networks into the EPC [62, 74, 76]. The standard accepts that the Wi-Fi network is as valid an access network as any other 3GPP radio access network. This acceptance enables operators to use the standards-based EPC components for integration and therefore helps ensure a good level of inter operability between different access types. For interworking between 3G and Wi-Fi, the 3GPP I-WLAN [75] standard defines the basic principles for managing the Wi-Fi networks in an integrated data offload scenario for the mobile network operator. 3GPP provides a solution to transfer data between the mobile device and the core network through a Wi-Fi access network. The underlying concept is to establish a controlled tunnel between the mobile device and a dedicated I-WLAN server in the core network in order to obtain access to operator subscribed content or public Internet. Furthermore, 3GPP release 10 proposed the ip flow mobility and Multi-Access PDN Connectivity for both EPC and I-WLAN architecture.5.1 shows the baseline architecture for multi-access PDN connectivity and IP flow mobility. The 3GPP Access Gateway acts as S-GW in case of trusted non-3GPPP access network such as WiMAx. In case of un-trusted non-3GPP access, like public

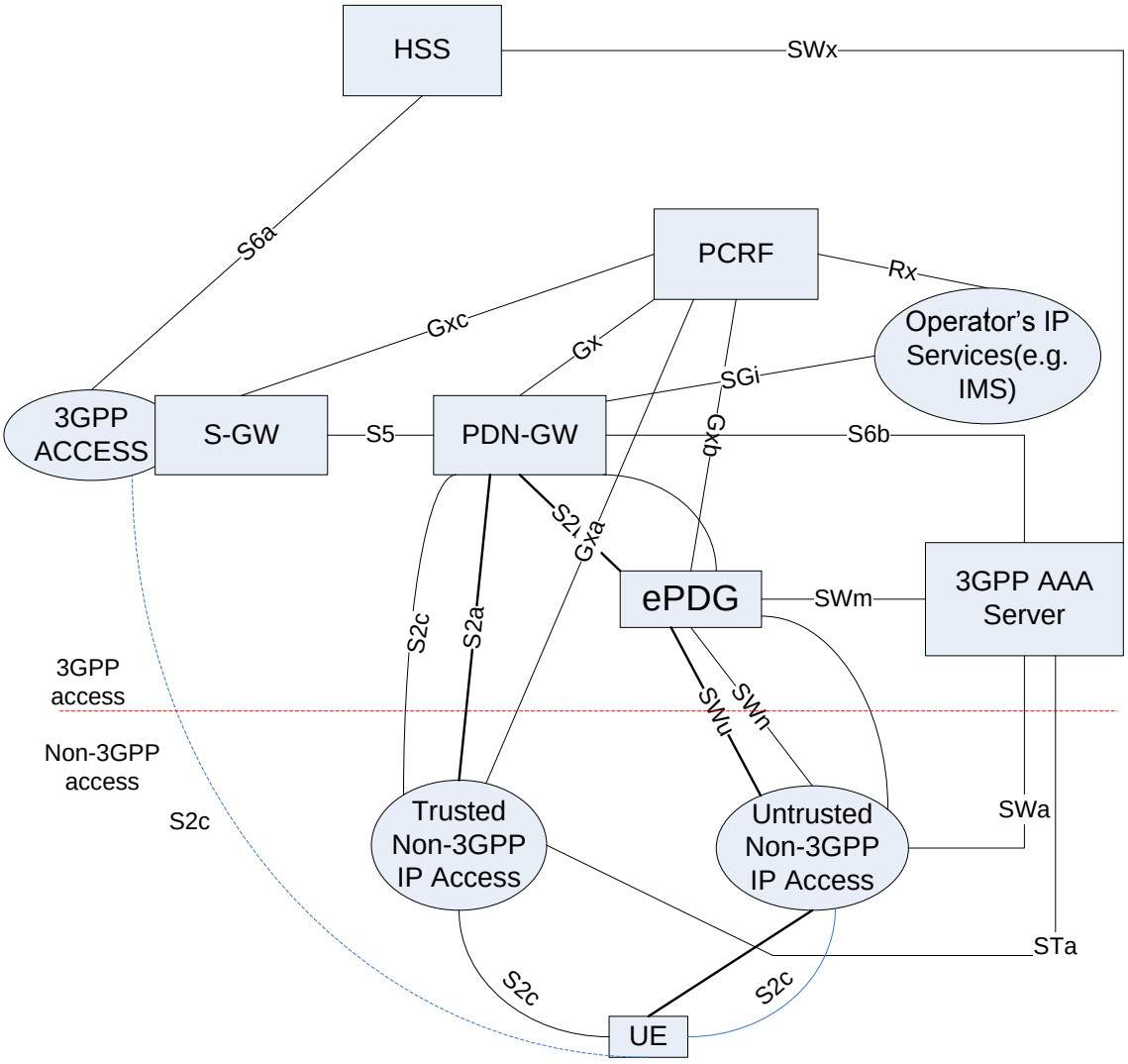

Figure 5.1: Non-Roaming 3GPP Arch. for Non-3GPP IP Access Integration into EPC using S5, S2a, S2b and S2c

Wi-Fi, the ePDG acts as S-GW to provide acceptable security by using IPSec between UE and the 3GPP network. The PDN-GW acts as anchor point for all uplink and downlink traffic to/from UE.

In this chapter, the focus is on trusted non-3GPP access network architecture i.e. Wi-Fi access network which is owned by the cellular operator. As it can be seen in Figure 5.2, *S2* and *S1* are the two interfaces that provide control and mobility support between non-3GPP access and P-GW, and forward the Wi-Fi traffic to the EPC. The *S2* interface provides mobility and control support between UE and P-GW over

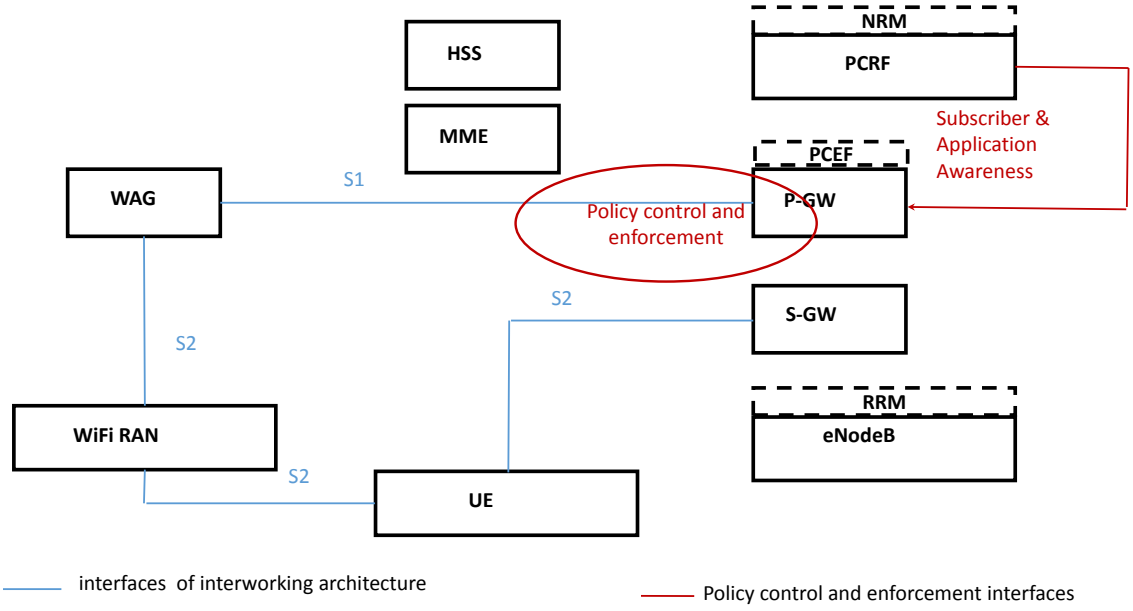

Figure 5.2: Trusted non-3GPP IP Access Integration into EPC

the non-3GPP access, while the *S1* interface, provides control and mobility support between trusted non-3GPP access and P-GW. The S-GW serves between LTE and Mobile access gateway (MAG) of Wi-Fi, and reports to the PCRF while also set the QoS parameters of bearers.

5.2 System Model

5.2.1 Policy Control

In the architecture of Figure 5.2, the Policy and Charging Enforcement Function (PCEF) in P-GW, enforces the policies and map service data flows to bearer to be mapped to the underlying transport network. The PCRF is LTE policy manager, which takes the operator policies, network information and user's profile stored in the HSS to make decisions based on the set of pre-defined rules and functions; QoS authorization, i.e. how to treat each traffic flow, is also performed by the PCRF. Today's policy controls are aware of users and applications while they are not aware of the congestion in the network [20]. While having fine-grain control on various entities in the mobile network is crucial for operators to allocate their resources, and to maintain and expand the network with low cost, it is also important to employ dynamic policies,

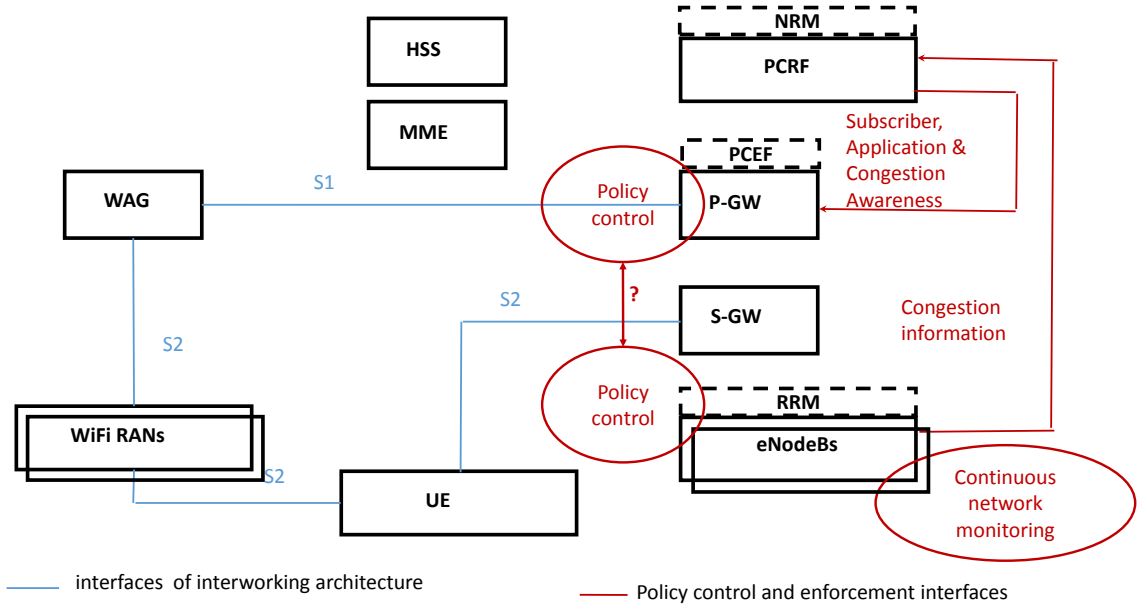

Figure 5.3: Coupling of RRM and NRM to optimize utilization of both radio and network resources

e.g. to manage the traffic and selectively offload the traffic between different access networks depending on the current network condition.

When performing traffic management, policy enforcement point can either be located at the core network or at the access network. At the core, enforcement is in the P-GW, and where the deep packet inspection (DPI) functions - similar to today's architecture in Figure 5.2. In this case, the service information does not have to travel at all, and the DPI engine can store subscriber information from the policy controller from which the congestion and location information can be estimated. The central solution as depicted here, requires simple integration of the DPI engine with the policy controller. On the other hand, the congestion information is extremely dynamic, i.e. it changes rapidly and by the time values sent from RAN are received by the P-GW, the information is not valid any more. Various studies show that 40% of the time bad decisions were made when outdated congestion information was used.

Moreover, estimation based on the current traffic is not simply possible for the DPI engine for various reasons. (1) The policy controller lacks a feedback mechanism. Simple questions such as whether the 1 Mbps for P2P is enough or it is being over penalized, can not be answered. (2) Throughput reduction is the only indication for the the policy controller, which can be either as a result of poor coverage or congestion.

Only the RAN can differentiate between these two causes of data rate drops. (3) Quite often the RAN is shared among different operators or perhaps the operator is using multiple access points. In these cases, policy controller can not see all traffic going through the congested cells. (4) Cell capacity depends on the coverage of the individual subscribers and varies even with weather conditions. There may be an additional reserved bandwidth for future bearers with guaranteed data rate, and other similar cases that affect the total cell capacity. In other words, the cell capacity varies with time and the policy controller receives no information about the capacity variations of the congested cells.

One alternative solution that can address the above challenges, is moving the policy enforcement point to the place where congestion occurs, so that there is no need to transfer this dynamic information anywhere. In this case, the scheduler is continuously prioritizing data packets and subscriber sessions. The scheduler has perfect knowledge about the location of the user, traffic and the real congestion conditions at that location. Then, policy controller will take service information from the DPI function and it will change the QoS parameters of the subscriber bearer, such as the traffic handling priority, the maximum bit rate, the guaranteed bit rate or the QoS Class Identifier (QCI). In the LTE, there is also the possibility to create a dedicated bearer for a specific traffic flow that requires a differentiated QoS treatment at the policy enforcement point. With these modifications, as it can be seen in Figuer 5.3, it is still not exactly clear where the policy control should be located.

5.2.2 SDN-Controller and Mobile Data Offloading

As discussed earlier, deriving real-time offloading policies for selectively offloading different services/applications based on the dynamic of network condition (Figure 5.3) could potentially be complex. On the other hand, a programmable interface similar to what SDN offers, facilitates offloading mechanism by providing an end-to-end communication between network elements and pushing corresponding forwarding rules to the local elements (i.e., eNodeB, and P-GW). In the considered architecture, the control-plane functionality of the gateways are decoupled and logically located as applications at the SDN-controller, while gateways run *local control agents*. The SDN-controller derives policies depending on the radio network conditions of the RRM application and the subscriber and application information of the PCRF application. This will allow operators to perform real-time traffic monitoring and provide per-subscriber QoS by programmable application modules in the SDN-controller and deriving forwarding rules accordingly. These policies and forwarding rules are periodically sent to the local

control agents (in the access network) to be forwarded to the UE. The LTE and Wi-Fi interworking architecture including a SDN-controller is depicted in Figure 5.4, where interactions with the local control agents are also shown. The two main parts of this architecture are as follows:

SDN-Controller

The SDN-controller in this architecture is an abstraction model that runs programmable applications modules such as RRM, and the PCRF. The PCRF application module is in possession of subscriber and application information, while RRM application collects radio access network condition such as traffic load and cell capacity. The SDN-controller combines the information of these two application module to derive a single set of policies and rules. These rules are sent periodically to the local control agents via the abstraction interface.

Local Control Agents

To address the challenges raised in Section 5.2.1, and also the scalability issue, local control agents is considered in the network gateways (P-GW, S-GW and WAG), and the RAN. These local control agents should perform measurement and some controlling actions which are authorized by the SDN-controller locally. For example, the agents that run on the gateways in this architecture, can measure QoS parameter such as delay and resource utilization and compare the traffic counters against the threshold and notify the SDN-controller in case of exceeding the threshold. To communicate back with the central controller, an interface similar to OpenFlow [77] is required at the local agents, which also allow them to perform simple control actions such as changing the weight or priority of a queue when the traffic counter exceeds a threshold.

5.3 Policy Derivation and Offloading Mechanism

5.3.1 Policy Derivation

The SDN-controller derives the offloading policy functions and rules by combining information from the RRM and PCRF applications. Radio network condition, defined by parameters such as wireless condition and traffic load, are measured frequently by the local control agents. Two offloading methods are detailed here, offline and online, where the offline method refers to the current architecture of fixed policy functions at

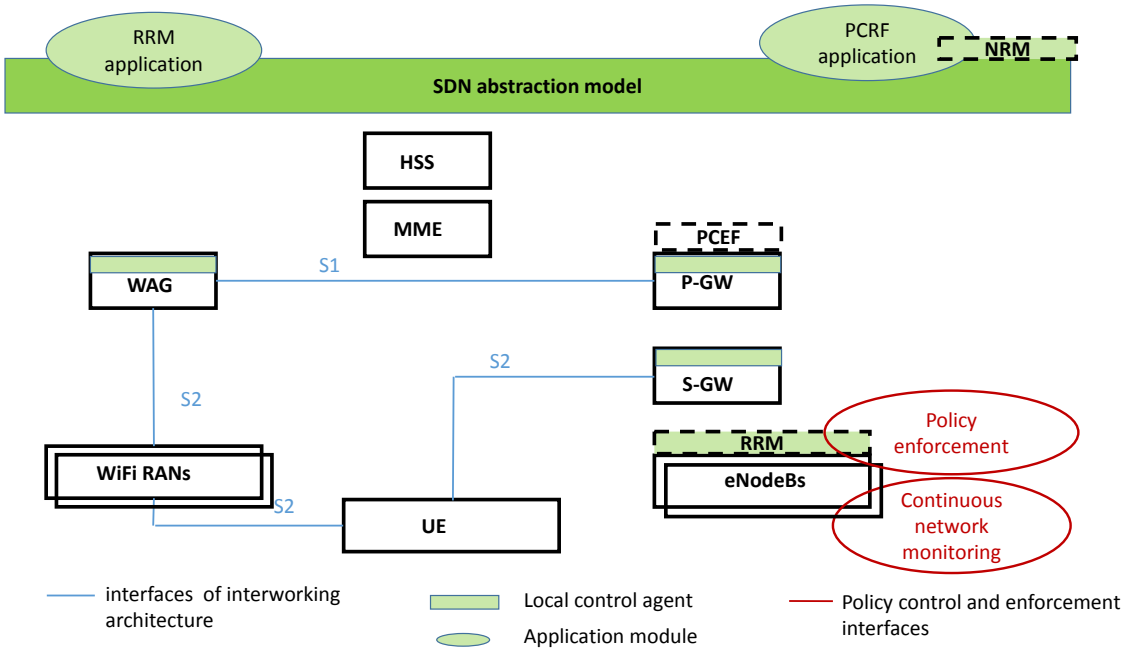

Figure 5.4: Coupling RRM and NRM via SDN-controller in the non-3GPP IP access integration into the EPC

the PCRF (Figure 5.2), and the online method refers to the policy decisions by the SDN-controller that are enforced via local control agents (Figure 5.4).

offline Method

In offline method, the PCRF identifies policies based on pre-defined set of functions and rules and depending on the list of subscribers and applications. These functions, set the *OffCap*, and the offered QoS by each network as fixed values and derive offloading policies accordingly. Threshold value of *OffCap* is the[] value at which offloading from LTE to Wi-Fi occurs, which is expressed in terms of percentage of the total capacity of LTE network (C_{lte}).

Given Λ is the LTE network utilization, $Q^0(wifi)$ is the QoS that can be offered by Wi-Fi network, then

$$\begin{cases} 1 - (\Lambda < \mathit{OffCap}) \text{ or } (\mathit{OffCap} \leq \Lambda < C_{lte} \text{ and } Q_i < Q^0\mathit{(min)}) \\ i \text{ is served by the LTE} \\ 2 - \Lambda \geq \mathit{OffCap} \text{ and } Q_i \geq Q^0\mathit{(wifi)} \\ i \text{ is offloaded to Wi-Fi} \\ 3 - \mathit{Otherwise} \quad i \text{ is dropped} \end{cases} \tag{5.1}$$

where Q_i represents a QoS parameter that has an upper bound, e.g. maximum delay that a flow can tolerate. The second part of line 1 in Equation (5.1), refers to the very delay sensitive applications (their delay requirement is less than a pre-defined minimum $Q^0\mathit{(min)}$), such as voice, which will be served by the LTE even after utilization is beyond *OffCap* (and as long as the LTE network is not over congested).

Online Method Scenario

In online method, network parameters are measured frequently by local control agents in the access network and are available at the RRM application. The SDN controller combine this information with the PCRF data and enforces policies by pushing the forwarding rules to the gateways. Combining these two application modules allow the offloading decision to be made based on the QoS requirement of that particular flow (user/application) as well as the offered QoS by each access network at any time.

As mentioned earlier, the offered QoS, e.g. the offered latency, by each network differs depending on the network condition and is noted by $Q^c(\text{lte})$ for the LTE and $Q^c(\text{wifi})$ for the Wi-Fi network. In this case, decision to either serve, offload, or drop the flow i will be as follows:

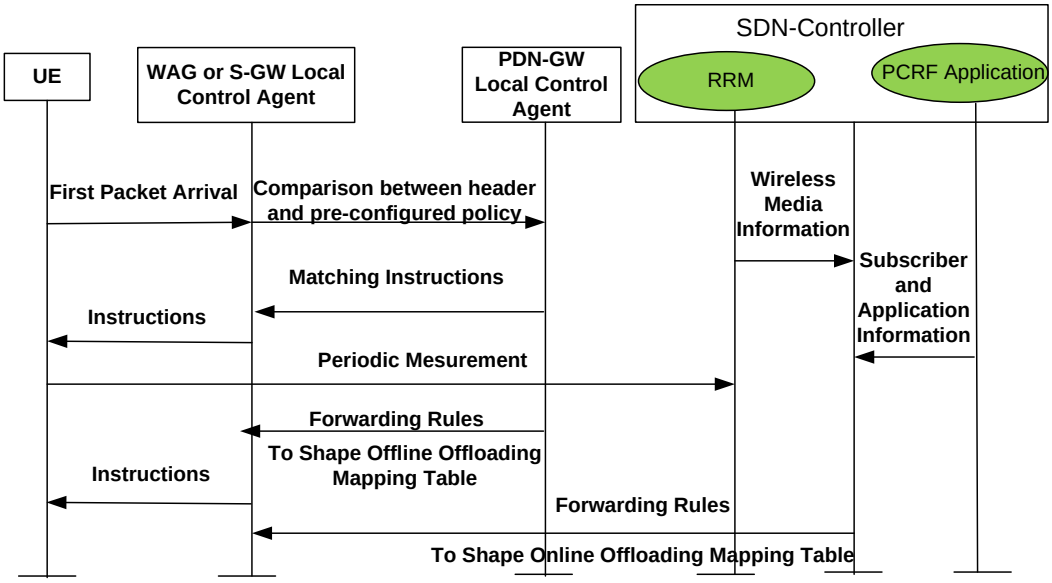

Figure 5.5: SDN enabled offloading procedure

$$\begin{cases} 1 - \Lambda < \mathit{OffCap} \\ i \text{ is served by the LTE} \\ 2 - \Lambda \geq \mathit{OffCap} \text{ and } Q_i \geq Q^c(\mathit{wifi}) \\ i \text{ is served by the Wi-Fi} \\ 3 - \mathit{OffCap} \leq \Lambda < C_{lte} \text{ and } Q^c(\mathit{wifi}) \geq Q_i \geq Q^c(\mathit{lte}) \\ i \text{ is served by the LTE} \\ 4 - \mathit{Otherwise} \quad i \text{ is dropped} \end{cases} \tag{5.2}$$

where Q^c is dynamically changing, depending on the current measurement from the network.

5.3.2 Offloading Mechanism

The policy derivation is performed in the SDN-controller, and the interworking between different modules is shown in Figure 5.5.

Initially UE transmits to the the either or both of the Wi-Fi and LTE access network gateway; upon arrival of the first packet of the flow, local control agents at the P-GW compares the header information with pre- configured policies and sends the match

offloading instruction to the access network gateway and the UE accordingly. After connection is established with the chosen access network, the following service data flows are mapped to bearers which match this policy. Then P-GW maps the bearers to the underlying transport layer.

The local control agents in the radio access network and the access gateways periodically collect the information such as utilization, and drop rate, from Wi-Fi and LTE access networks and report these measurements to the SDN-controller. The SDN-controller will then combine these with the PCRF information, derive the policy functions and forwarding rules, and push them back to the local control agents. Hence, the local control agent will shape an offloading mapping table based on these rules. The updating frequency of these rules and functions can be set according to the frequency of change in the network conditions as well as the availability of communication resources between the controller and the access network.

Hence, the proposed architecture offers (1) **programmable** offloading policies by converting PCRF and RRM to software modules; (2) **simplicity** by abstraction of the network view and simplifying the communication between network elements, (3) enhanced **efficiency** by allowing more up to date and precise congestion information to reflect on the offloading policies. In the next section, we show how such an architecture can enhance the network performance in quantifiable metrics.

5.4 performance Evaluation

The proposed architecture is used to offload the traffic based on the PCRF module policy derivation. The main focus of this thesis is on the programmable offloading policies, that is managed by a logically centralized SDN-controller (based on the architecture suggested in [78]), while issues such as mobility management and authentication are out of the scope of this work.

5.4.1 Simulation Model

The simulation model considers a single cell served by an omni-directional antenna, and OFDMA downlink (general 4G/LTE model) with the total bandwidth of 20 MHz and the downlink data rate of 100 Mbit/s. The cell also covered by a number of randomly distributed Wi-Fi access networks, which their coverage area does not overlap each other, i.e. each mobile user can only get service via a single Wi-Fi access point at any location.

Table 5.1: Performance Requirement by Service Category

Service ID	Service Type	Rate (kbps)	Delay (ms)
1	VoIP	21-64	50-100
2	Video	500-700	200-300
3	interactive gaming	300-600	100-300
4	Internet	50-600	300-600
5	Peer-to-Peer	700-1000	300-600
6	Business Services	600-800	50-100

For the channel model, a propagation loss model has been considered [58]. The contributing factors affecting the received power at the UEs in such a model are free space path loss and shadowing. The free path loss between eNodeB and user is determined using standard radio propagation models which consider the loss, L, as a function of the distance between eNodeB and the user in Km, d, on the form as defined by the equation given below.

$$P_L = 128.1 + 37.6 * \log d \tag{5.3}$$

Shadowing is modeled as a log-normal random variable with zero mean and standard deviation of 8dB.

It is assumed further that there are enough Wi-Fi resources to maintain the current Wi-Fi data rate; hence data rate that Wi-Fi access point can offer is independent of its traffic load in the network. As Wi-Fi and LTE networks operate on different frequency bands, there is no resource partitioning or interference between these access networks. Interference between adjacent Wi-Fi access points (relying on the Wi-Fi planning strategy), load balancing across them, and power control across overlapping access networks are not considered in this model.

On the other note, all services in the LTE are provided as packet services including voice services. Hence, real-time and non real-time services are multiplexed over the air interface and core network. For modeling traffic, six categories of services as voice over IP, video, gaming and interactive, Internet, business services and peer-to-peer are defined. Each service category $k \in \{1, ..., 6\}$ is characterized with uniformly distributed random delay requirement of d_k and data rate of b_k. The boundaries for each value depending on service type are detailed in Table 5.1.

The proposed approach is general and it is independent from the model used for

describing incoming data and stochastic flow modeling is not needed. A flow is defined as a signal modeling the bit rates produced by application layer. In the system, it is assumed that N active traffic flows share the wireless channel. Packets waiting for transmission are sorted in a queue associated to the user buffer for each type of traffic. An infinite buffers is assumed for each user to carry one of the services mentioned above each time [].

A set of n active flows is denoted by F i.e. $f_i \in F$, where f_i is identified by d_i and b_i of the related service category. Current traffic load of cellular network is uniformly distributed random variables in the range of [10 Mbps, 50 Mbps]. Users' throughput is computed depending on the current traffic load, and the service type of the incoming flow i and physical channel. For example, flow i_1 that is of service type k_1, will receive $\alpha \times \beta \times b_{k_1}$ as their throughput, where α shows how busy the network is. In this work, the network can handle up to 100 Mbps, and thus α varies in the range of 90% and 50%, depending on the network traffic load, β is a coefficient to show the channel condition and it is estimated based on the SNR and CQI curves described in 3
Change in SNR means change in the spectral efficiency which is respectively affect the throughput. Similarly for the Wi-Fi network, utilization is a uniformly distributed random value between 1 Mbps and 40 Mbps. Each Wi-Fi access network is considered to be a 802.11n with the maximum data rate of 60 Mbps. In consecutive rounds of simulation, number of active flows are increased from 180 to 270.

Mobile users are distributed uniformly over the cell with radius 1 Km. To uniformly distribute the mobile users, a cartesian coordinate system is considered where x and y coordinates are chosen uniformly.

5.4.2 Simulation Scenarios and Numerical Results

The examined metrics here include **dropping rate** and **offloaded traffic rate**. These two figures of merit are defined as follows: (1) dropping rate is the ratio of dropped traffic to the overloaded traffic (2) offloaded traffic rate is the ratio of offloaded traffic to the overload traffic on the LTE. Simulations run separately for the online and off-line methods, and based on their related constraints.

In both online and offline method *OffCap* is set to 60% of i.e. the threshold at which offloading from LTE to Wi-Fi is triggered. In the offline method, offered delay by Wi-Fi network is set to the fixed values of 200 ms (Q^0*(min)*)independent of service category of the flow, and LTE network admits only delay sensitive flows which they have delay requirement below 100 ms (Q^0*(min)*). When offloading triggers, if Wi-Fi

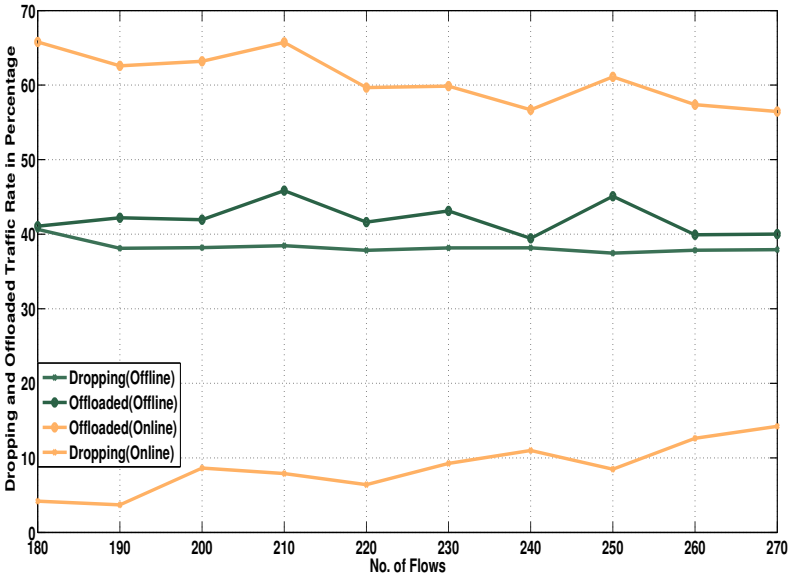

Figure 5.6: Offloaded Traffic Rate, and Dropping Rate in percentage Vs No. of Flows for Online and Offline Policy Derivation Methods.

satisfies delay, and data rate requirements of the flow, then the flow is offloaded from LTE to Wi-Fi; otherwise it is dropped.

In online method, offloading decisions are made depending on the delay and bandwidth requirements of each flow, and the current condition of Wi-Fi and LTE network (collected by measurements as described in Section 5.3). When offloading triggers, if Wi-Fi network can accommodate delay and data rate requirement of flow i, then this flow is offloaded, otherwise if its delay and data rate requirement can be satisfied by LTE, it will be served by LTE. Finally LTE network will drop this flow if it can not accommodate its delay requirement or if by adding that flow to the network, traffic load goes beyond the maximum network capacity.

Scenario One: Same Distance from eNodB for all Users

In this scenario, it is assumed that users are stationary and their distances from the access network in each simulation run are constant, therefore the path loss coefficient constant. Therefore, users' throughput is not affected by its distance to the base station. The results in Figure 5.6 show the offloaded traffic rate and the dropping rate for online and offline methods. The online method shows an average 15% to 35% higher offloaded traffic rate and 10%-20% lower dropping rate comparing to offline method. In

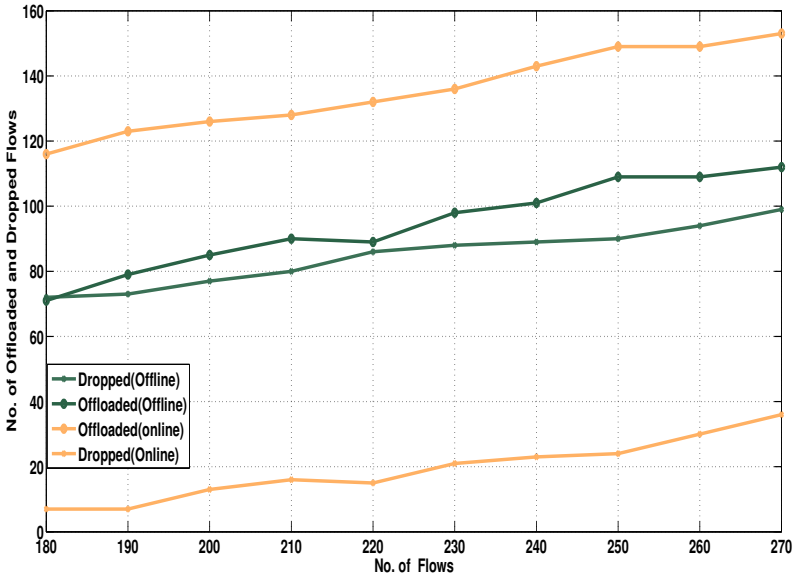

Figure 5.7: No. of Offloaded and Dropped Flows vs No. of Flows for Online and Offline Policy Derivation Method

offline method, as a consequence of fixed offered delay in Wi-Fi, less flows are offloaded to Wi-Fi, which results in higher dropping rate. The online method, considering the current conditions of the Wi-Fi and LTE, more flows are offloaded on Wi-Fi network, or more flows are served by LTE and less flows are dropped. It should be noted that, in each round of simulation on average 40% of the traffic is dropped in offline method. This is due to offline offloading policies which does not consider the real-time condition of the access networks. When the number of flows increases, the traffic load generated increases accordingly, while the Wi-Fi capacity and LTE capacity remains the same which results in the increase in the dropping rate of online method.

Figure 5.7 shows offloaded and dropped flows for both online and offline method. In online method comparing to offline method on average 20 more flows are offloaded to Wi-Fi network and 50 less flows are dropped.

significant improvement in system performance through online offloading mechanism.

The presented results here, shows that applying online method policies through SDN-controller considering the current network condition, and perform offloading based on programmable and flexible policies rather than fixed policies improves the network performance significantly.

Scenario Two: Changing the Users' Distance from eNodeB

In this scenario, low mobility users are assumed. In each round of simulation, the distances of users from the eNodeB are changed. Each user moves from its current location within the range of 10 m to 70 m. Therefore, the user mobility pattern affects the throughput due to the change in channel condition.

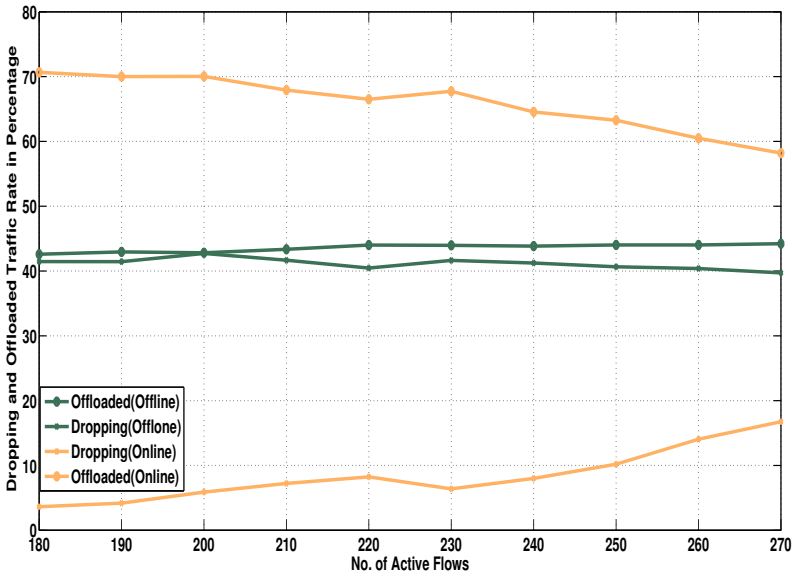

Figure 5.8: Offloaded Traffic Rate, and Dropping Rate in percentage Vs No. of Flows for Online and Offline Policy Derivation Methods.

5.5 Concluding Marks

Mobile data offloading has already become a key solution to address the explosive data traffic demand over mobile networks. In this chapter, a programmable policy function derivation framework through applying an abstract of SDN in the mobile backhaul is proposed. This framework consider the real-time network condition measurement to derive the offloading policies and efficiently accommodate the traffic to the LTE or Wi-Fi network. The simulation results shows that this mechanism can achieve up to 35% increase in offloaded traffic from LTE to Wi-Fi while it results in 15% less dropping rate. On the other note, the significant increase in the number of offloaded flows and decrease in the dropped flows points to network performance improvement. In addition to those, the proposed architecture offers programmable policy functions by converting

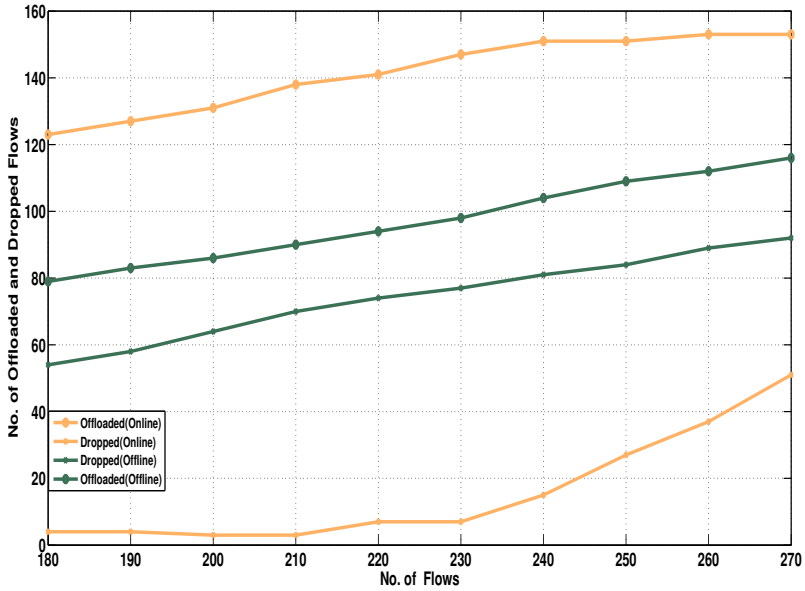

Figure 5.9: No. of Offloaded and Dropped Flows vs No. of Flows for Online and Offline Policy Derivation Method

PCRF to a software module, and further evolvability by running an OS-like controller in which changes in the network can be easily configured.

Chapter 6

Conclusions and Future Research

This thesis discusses my views in the design of a smart QoS-aware data offloading. The objective of this design is to improve the performance of the end-to-end QoE for users, using enhanced architectures and mechanisms that are compatible with the existing architectures and require no modification. To this end, three QoS-aware offloading mechanisms are proposed:QoS-aware resource allocation and network selection mechanisms; policy based offloading with an autonomic networking concept; and programmable policies for mobile data offloading using the abstraction of SDN and an openswitch interface. Performance of the proposed schemes has been investigated by extensive simulations. Various figures of merit, such as the end-to-end throughput, and average delay for different type of flows, have been explored to present the efficiency of the proposed algorithms in a heterogenous network consisting of LTE and Wi-Fi.
The first proposed mechanism aims to maximize throughput while maintaining the QoS of the flows. In this mechanism, the delay experienced by the flows is examined in the first TTI or frame, prior to transmission, and the decision can be made either to offload the flows to Wi-Fi, or to serve them in the cellular network or drop them, if they experienced delay greater than their delay constraint in both networks. Utilizing the QoS requirement of each flow, such as delay and throughput, a smart offloading framework has been detailed to selectively and fairly allocate the resources among users. The real-time solution of the proposed problem has been discussed using an IP-flow mobility and QoS aware scheduling in the MAC scheduler, while numerical observations have confirmed the increase in the throughput and number of flows served on one side, and the subsequent decrease in the average queuing delay. Thorough simulation studies, using Matlab simulation under various channel conditions and for a mixed class of services, have revealed that system performance is improved significantly.

Secondly, state- of- the- art, policy-based decision making is presented. In the

proposed policy- based offloading framework, offloading decisions a re based on a set of rules that consider both real-time network conditions and operator objectives and strategies. Three different approaches have been investigated. The network centric approach considers network and environmental conditions in the decision making, and guarantees a better stability and robustness in the system. However, it introduces a higher signalling load. A user centric approach alleviates the signalling overload and introduces better adaptation to the user profile, thus enhancing the QoS. However, this violates the overall system stability. A better approach is to employ hybrid decision making which is a trade-off between the system stability and QoS enhancement for the user. A hybrid policy is based on principles of autonomic networking and is introduced where decision are shared between user and network. The performance of the proposed framework is evaluated through a cost-function approach. Detailed simulations examine the performance metric, such as offloading efficiency (OE) and a blocking ratio(BR), and show significant improvement in the system performance.

Finally, a programmable, policy function derivation framework through applying an abstract of SDN in the mobile backhaul, is proposed. The proposed mechanism considers the real-time network conditions to derive the offloading policies and efficiently accommodate the traffic in both LTE and Wi-Fi networks. Numerical results prove that the proposed approach can significantly improve the drop rate of the incoming traffic by using more real-time and dynamic decisions for offloading. In this way, a significant increase in the number of offloaded flows and a decrease in the dropped flows point to network performance improvement. In addition to these initiatives, the proposed architecture offers programmable policy functions by converting PCRF to a software module, and further opportunities for development ,by running an OS-like controller in which changes in the network can be easily configured. The architecture also offers enhanced efficiency by providing more up to date and precise congestion information for reflection on the offloading policies.

6.1 Avenues of Future Research

In this section, some of the most interesting new topics among many others emerging technologies that can be continued by the research is presented.

6.1.1 Opportunistic small cells and 3G/4GWi-Fi interworking

This thesis mainly focused on the network and user performance in heterogenous network consisting of Wi-Fi access network. However, Following the deployment of HetNets, industry is already working on future enhancements. One such example is opportunistic small cells which dynamically switch ON and OFF based on the presence or absence of traffic. This would not only reduce interference, but also save energy.

6.1.2 Offloading Decision in a Framework of HetNet and Mobile Cloud Computing

Further research in this area could be on a framework proposed in [79] which is based on QoS and power negotiation between the offloading decision module and the wireless HetNet. Offloading decision computation can be investigated in a tight coupling of mobile cloud computing applications and wireless heterogenous network. On the other hand, wireless HetNet should perform RRM functions such as admission control and resource allocation to provide QoS guarantee. Some of the Open research problem in this area are presented in [80].

6.1.3 Software Defined Cloud Networking (SDCN)

This thesis, introduced the programmable offloading policies through applying abstraction of SDN. Software Defined Networking (SDCN) has become an unstoppable force by combining the principles of cloud computing, such as automation, self service provisioning, and linear scaling of both performance and economics that can deliver network virtualization, custom programmability, and simplified architectures. This combination creates a best obtainable software foundation in order to maximize the value of the network to both the users and service providers. Therefore, new architectures can simplify management and provisioning, speed up service delivery, lower costs and create opportunities for competitive differentiation, while putting control and visibility back in the hands of the network operators by dynamically deriving policies and then applying these to the network. SDN can be seen as a complementary technology to virtualization and is potentially well suited for a network-enabled cloud and improved network resource utilization at the link level.

References

[1] "VNI Global IP Traffic Forecast, 2012-2017," Cisco White paper, May 2013. 16, 23

[2] K. Samdanis, T. Taleb, and S. Schmid, "Traffic Offload Enhancements for eUTRAN," *Communications Surveys Tutorials, IEEE*, vol. 14, no. 3, pp. 884–896, 2012. 21

[3] "3G TS 23.261 Version 10.1.0 : Technical Specification Group Services and System Aspects; IP flow mobility and seamless Wireless Local Area Network (WLAN) offload," Sept. 2010.

[4] "3G TR 23.829 Version 10.0.0: Technical Specification Group Services and System Aspects; Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)," Mar. 2011. 21

[5] "TR 36.913, V8.0.0, Requirements for Further Advancements for E-UTRA (LTE-Advanced), 3rd Generation Partnership Project, Technical Specification Group Radio Access Network," Jun. 25

[6] K. Zheng, B. Fan, J. Liu, Y. Lin, and W. Wang, "Interference coordination for OFDM-based multihop LTE-advanced networks," *Wireless Communications, IEEE*, vol. 18, 2011. 26

[7] K. Zheng, Y. Wang, W. Wang, M. Dohler, and J. Wang, "Energy-efficient wireless in-home: the need for interference-controlled femtocells," *Wireless Communications, IEEE*, vol. 18, no. 6, 2011. 26

[8] "IEEE Standard for Local and Metropolitan Area Networks: Air Interface for Fixed Broadband Wireless Access Systems: Medium Access Control Modifications and Additional Physical Layer Specifications for 2-11 GHz," *IEEE Std 802.16a-2003 (Amendment to IEEE Std 802.16-2001)*, 2003. 26

[9] "IEEE Standard for Local and Metropolitan Area Networks: Air Interface for Fixed Broadband Wireless Access Systems - Medium Access Control Modifications and Additional Physical Layer Specifications for 2-11 GHz Incorporated into 802-16-2004," *IEEE Std P802.16a/D7.0*, 2002. 26

[10] "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1," *IEEE Std 802.16e-2005 and IEEE Std 802.16-2004/Cor 1-2005 (Amendment and Corrigendum to IEEE Std 802.16-2004)*, 2006. 27

[11] "Femtocells - Natural Solution for Offload," Femto Forum White Paper, Jun. 2011. 31

[12] "3GPP TS 23.859 Version 12.0.1: Technical Specification Group Services and System Aspects; Local IP access (LIPA) mobility and Selected IP Traffic Offload (SIPTO) at the local network," Apr. 2013. 32

[13] Y. Choi, H. W. Ji, J. yoon Park, H. chul Kim, and J. Silvester, "A 3W network strategy for mobile data traffic offloading," *Communications Magazine, IEEE*, vol. 49, no. 10, pp. 118–123, 2011. 33

[14] B. Han, P. Hui, V. A. Kumar, M. V. Marathe, G. Pei, and A. Srinivasan, "Cellular traffic offloading through opportunistic communications: a case study," in *Proceedings of the 5th ACM workshop on Challenged networks*, ser. CHANTS '10, 2010, pp. 31–38. 33, 44

[15] L. Xiaofeng, H. Pan, and P. Lio, "Offloading mobile data from cellular networks through peer-to-peer WiFi communication: A subscribe-and-send architecture," *Communications, China*, vol. 10, no. 6, 2013.

[16] T. Han, N. Ansari, M. Wu, and H. Yu, "On Accelerating Content Delivery in Mobile Networks," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 3, 2013. 33

[17] A. De La Oliva, C. Bernardos, M. Calderon, T. Melia, and J. Zuniga, "IP Flow Mobility: Smart Traffic Offload for Future Wireless Networks," *IEEE Commun. Mag.*, vol. 49, no. 10, pp. 124–132, 2011. 34, 45

[18] "Media Optimization for Mobile Networks, white paper by Openwave," Sept. 2011. 36

[19] "Quality of Service (QoS) and Policy MAnagement in Mobile Data Networks," IXIA, July 2011. 38

[20] "3GPP TS 23.203, Policy and Charging Control Architecture," Mar. 2011. 38, 41, 85

[21] I. Sesia, S. Toufik and M. Baker, "LTE - The UMTS Long Term Evolution: From Theory to Practice," John Wiley and Sons, Ltd, Chichester, UK, 2009. 39, 47, 62

[22] "Distributed Decisios: Heirarchical network Policy Control," Sandvine Intelligent Broadband Networks, 2011. 42

[23] "Cisco Enterprise Policy Manager," Cisco, Feb. 2008. 42

[24] S.-I. Sou and C.-S. Lin, "SPR proxy mechanism for 3GPP Policy and Charging Control System," *Comput. Netw.*, vol. 55, no. 17, Dec. 2011. 43

[25] E. Gustafsson and A. Jonsson, "Always Best Connected," *Wireless Communications, IEEE*, vol. 10, no. 1, 2003. 44

[26] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3g using wifi," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, ser. MobiSys '10, 2010. 44

[27] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile Data Offloading: How Much Can WiFi Deliver?" *IEEE/ACM Trans. on Net.*, vol. 21, no. 2, pp. 536–550, Apr. 2013. 44, 77, 78

[28] H. Zhou, K. Sparks, N. Gopalakrishnan, P. Monogioudis, F. Dominique, P. Busschbach, and J. Seymour, "Deprioritization of heavy users in wireless networks," *Communications Magazine, IEEE*, vol. 49, no. 10, pp. 110–117, 2011. 44

[29] H. Izumikawa and J. Katto, "RoCNet: Spatial mobile data offload with user-behavior prediction through delay tolerant networks," in *Wireless Communications and Networking Conference (WCNC), 2013 IEEE*, 2013. 45

[30] S.-I. Sou, "Mobile Data Offloading With Policy and Charging Control in 3GPP Core Network," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 7, 2013. 45

[31] Y. Choi, H. W. Ji, J. yoon Park, H. chul Kim, and J. Silvester, "A 3W network strategy for mobile data traffic offloading," *Communications Magazine, IEEE*, vol. 49, no. 10, 2011. 45

[32] M. Bennis, M. Simsek, A. Czylwik, W. Saad, S. Valentin, and M. Debbah, "When cellular meets WiFi in wireless small cell networks," *Communications Magazine, IEEE*, vol. 51, no. 6, 2013. 45

[33] S. Paris, F. Martignon, I. Filippini, and L. Chen, "A bandwidth trading marketplace for mobile data offloading," in *INFOCOM, 2013 Proceedings IEEE*, 2013, pp. 430–434. 45

[34] D. Holma and D. A. Toskala, "LTE for UMTS-OFDMA and SC-FDMA Based Radio Access," Wiley Publishing, 2009. 45, 62

[35] "3GPP TS 23.107 Version 11.0.0: Technical Specification , Digital cellular telecommunications system (Phase 2+); Quality of Service (QoS) Concept and Architecture," Nov. 2012. 47

[36] "3G/Wi-Fi seamless Offload," Qualcomm White Paper, March 2010. 48

[37] G.Tsirtsis and H.Soliman, "IETF RFC5555, Network Working Group, Mobile IPv6 Support for Dual Stack Hosts and Routers," Jun. 2009. 48

[38] "LTE MAC Scheduler & Radio Bearer QoS," Roke Manor Research Limited, 2011. 52

[39] "LTE eNodeB MAC Scheduler Interface," Roke Manor Research Limited, 2009. 53

[40] "LTE MAC Scheduler & Radio Resource Scheduling," Roke Manor Research Limited, 2011. 53

[41] Y. J. Zhang and K. Letaief, "Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems," *Wireless Communications, IEEE Transactions on*, vol. 3, no. 5, pp. 1566–1575, 2004. 54

[42] J. Jang and K.-B. Lee, "Transmit power adaptation for multiuser OFDM systems," *Selected Areas in Communications, IEEE Journal on*, vol. 21, no. 2, pp. 171–178, 2003. 54

[43] B. Sadiq, S. J. Baek, and G. de Veciana, "Delay-Optimal Opportunistic Scheduling and Approximations: The Log Rule," *Networking, IEEE/ACM Transactions on*, vol. 19, no. 2, pp. 405–418, 2011. 54

[44] S. Shakkottai and A. L. Stolyar, "Scheduling for Multiple Flows Sharing a Time-Varying Channel: The Exponential Rule," *American Mathematical Society Translations, Series*, vol. 2, p. 2002, 2000.

[45] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates," *Probab. Eng. Inf. Sci.*, vol. 18, no. 2, pp. 191–217, Apr. 2004. 54

[46] S. Schwarz, C. Mehlfuhrer, and M. Rupp, "Low complexity approximate maximum throughput scheduling for LTE," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*, 2010, pp. 1563–1569. 54

[47] S.-B. Lee, I. Pefkianakis, A. Meyerson, S. Xu, and S. Lu, "Proportional Fair Frequency-Domain Packet Scheduling for 3GPP LTE Uplink," in *INFOCOM 2009, IEEE*, 2009, pp. 2611–2615.

[48] L. Le, E. Hossain, and A. Alfa, "Service differentiation in multirate wireless networks with weighted round-robin scheduling and arq-based error control," *Communications, IEEE Transactions on*, vol. 54, no. 2, pp. 208–215, 2006. 54

[49] M. Wemersson, S. Wanstedt, and P. Synnergren, "Effects of QoS Scheduling Strategies on Performance of Mixed Services over LTE," in *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*, 2007, pp. 1–5. 54

[50] M. Gidlund and J.-C. Laneri, "Scheduling Algorithms for 3GPP Long-Term Evolution Systems: From a Quality of Service Perspective," in *Spread Spectrum Techniques and Applications, 2008. ISSSTA '08. IEEE 10th International Symposium on*, 2008, pp. 114–117. 54

[51] H. Lei, M. Yu, A. Zhao, Y. Chang, and D. Yang, "Adaptive Connection Admission Control Algorithm for LTE Systems," in *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*, 2008, pp. 2336–2340. 54

[52] B. Sadiq, S. J. Baek, and G. de Veciana, "Delay-Optimal Opportunistic Scheduling and Approximations: The Log Rule," *Networking, IEEE/ACM Transactions on*, vol. 19, no. 2, pp. 405–418, 2011. 54

[53] G. Piro, L. Grieco, G. Boggia, R. Fortuna, and P. Camarda, "Two-Level Downlink Scheduling for Real-Time Multimedia Services in LTE Networks," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, pp. 1052–1065, 2011. 55

[54] "A Surevey of Scheduling Theory in Wireless Data networks." 55

[55] S. J. Baek and G. de Veciana, "Opportunistic Feedback and Scheduling to Reduce Packet Delays in Heterogeneous Wireless Systems," *Vehicular Technology, IEEE Transactions on*, vol. 61, no. 7, 2012. 55

[56] Boyd, Stephen and Vandenberghe, Lieven, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004. 59

[57] T. Camp, J. Boleng, and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research," *WIRELESS COMMUNICATIONS & MOBILE COMPUTING (WCMC): SPECIAL ISSUE ON MOBILE AD HOC NETWORKING: RESEARCH, TRENDS AND APPLICATIONS*, vol. 2, pp. 483–502, 2002. 61

[58] "3GPP, Tch, Specif. Group Radio Access Network; Physical layer aspect for evolved Universal Terrestrial Radio Access(UTRA), 3GPP TS 25.814 (Release 7)." 61, 93

[59] G. Piro, L. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE Cellular Systems: An Open-Source Framework," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 2, pp. 498–513, 2011. 62

[60] S. Dahlman, J.S.E.Parkvall and P. Beming, "3G Evolution HSPA and LTE for Mobile Broadband," New York, Academic Press, USA, 2008. 62

[61] S. Na and S. Yoo, "Allowable Propagation Delay for VoIP Calls of Acceptable Quality," in *Proceedings of the First International Workshop on Advanced Internet Services and Applications*, ser. AISA '02. London, UK, UK: Springer-Verlag, 2002, pp. 47–56. 62

[62] "3GPP TS 23.302, Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks," Apr. 2011. 62, 83

[63] P. Leaves, K. Moessner, R. Tafazolli, D. Grandblaise, D. Bourse, R. Tonjes, and M. Breveglieri, "Dynamic spectrum allocation in composite reconfigurable wireless networks," *Communications Magazine, IEEE*, vol. 42, no. 5, pp. 72–81, 2004. 68

[64] J. Mitola, "Cognitive radio," Ph.D. dissertation, KTH, Teleinformatics, 2000. 69, 74

[65] W. Shen and Q.-A. Zeng, "Cost-Function-Based Network Selection Strategy in Integrated Wireless and Mobile Networks," *Vehicular Technology, IEEE Transactions on*, vol. 57, no. 6, pp. 3778–3788, 2008. 70

[66] J. Kephart and D. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003. 72

[67] A. Mihailovic, G. Nguengang, J. Borgel, and N. Alonistioti, "Building Knowledge Lifecycle and Situation Awareness in Self-Managed Cognitive Future Internet Networks," in *Emerging Network Intelligence, 2009 First International Conference on*, 2009. 72

[68] A. Mihailovic, G. Nguengang, A. Kousaridas, M. Israel, V. Conan, I. Chochliouros, M. Belesioti, T. Raptis, D. Wagner, J. Moedeker, V. Gazis, R. Schaffer, B. Grabner, and N. Alonistioti, "An approach for designing cognitive self-managed Future Internet," in *Future Network and Mobile Summit, 2010*, 2010. 72

[69] "A 3G/LTE Wi-Fi Offload Framework: Connectivity Engine (CnE) to Manage Inter-System Radio Connections and Applications Qualcomm," Qualcomm White Paper, Jun. 2011. 73

[70] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "A simple analytical model for the energy-efficient activation of access points in dense WLANs," in *Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking*, ser. e-Energy '10. ACM, 2010. 75

[71] A. Mahanti, C. Williamson, and M. Arlitt, "Remote analysis of a distributed WLAN using passive wireless-side measurement," *Perform. Eval.*, vol. 64, no. 9-12, Oct. 2007. 75

[72] A. Aijaz, O. Holland, P. Pangalos, and H. Aghvami, "Energy savings for cellular access network through Wi-Fi offloading," in *Communications (ICC), 2012 IEEE International Conference on*, 2012. 77

[73] "A New Traffic Model for Current UserWeb Browsing Behavior," white paper, Intel Corporation, Sep. 2007. 77

[74] "3GPP TS 23.402 V10.7.0: Architecture enhancements for non-3GPP accesses," http://www.cellular-news.com/story/46917.php, Mar. 2013. 83

[75] "3GPP TS 23.234, 3GPP system to Wireless Local Area Network (WLAN) interworking;System description," Nov. 2012. 83

[76] "Architecture for Mobile Data Offload over WiFi Access Networks," Cisco White Paper, 2012. 83

[77] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: enabling innovation in campus networks," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, mar 2008. 88

[78] L. E. Li, Z. M. Mao, and J. Rexford, "Toward Software-Defined Cellular Networks," in *EWSDN*, Oct. 2012. 92

[79] L. Lei, Z. Zhong, K. Zheng, J. Chen, and H. Meng, "Challenges on Wireless Heterogeneous Networks for Mobile Cloud Computing," *Wireless Communications, IEEE*, vol. 20, 2013. 101

[80] L. Gkatzikis and I. Koutsopoulos, "Migrate or not? exploiting dynamic task migration in mobile cloud computing systems," *Wireless Communications, IEEE*, vol. 20, no. 3, 2013. 101